Mutational and structural analyses of Tomato curly stunt virus pathogenicity determinants

by

Alexander Michal Zwolinski

(1818803)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Prof M. E. C. Rey

Co-supervisor: Dr I. A. Achilonu

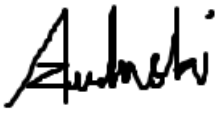
May 2022

Declaration

I, Alexander Zwolinski (1818803), am a student registered for the degree Doctor of Philosophy (MCBG 9002) in the academic year 2021.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the Thesis HAS NOT been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: 30 May, 2022.

Acknowledgements

What a fantastic journey it has been. I have met many fascinating people and learnt some valuable lessons along the way. I am eternally grateful to the following:

Prof Chrissie Rey, my supervisor, for her resolute enthusiasm even in the face of seemingly impossible challenges. She has inspired me to fully immerse myself in this incredible field. I am so grateful to her for encouraging me to attend workshops, collaborate with others and approach every single task with curiosity, passion, and determination. Cheers to you Prof!

Dr Lindy Esterhuizen for entrusting me with her precious ToCSV infectious clones – I have become so fond of them; I hope that she will appreciate the findings of this study especially since she provided key advice.

Frances Ayres, for teaching me nearly everything I know about protein work – she has truly been an incredible support and I will always remember our adventure-packed trip to India. I deeply admire and respect her.

Alison Brigden, my lab mate through thick and thin. She has been one of the most positive and uplifting influences throughout. It has been an absolute pleasure to work with her and I am deeply indebted to her. She has carried me through many tough times with a laugh, a crazy dance and a plenty of koekies!

Dr Ike Achilonu, for encouraging me to tackle recombinant expression of a particularly stubborn protein. It has been challenging, but I thank him for persisting and offering me advice along the way.

My colleagues in MCB – Dr Imah Mwaba, Dr Farhahna Allie, Dr Helen Walsh, Dr Patience Chatukuta, Nikita, Warren, Nick and Kate for all the laughs and intense discussions! I have many fond memories of our time together.

The friends I have made from other departments – a special thank you to Gaby for the wine, Zaahida for the coffee, Dean for the laughs, Jessica for the real talk, and Lauren for spinning the records.

Dr Angela Botes for kindly letting me use her lab facilities and for all the support, it has also been such a pleasure to get to know your students!

Dr Vanessa Meyer for always offering her encouragement, Michael Tobin and Lucretia Brink for assisting with growth facilities, Wilhelmina for her cheerfulness.

Prof Supriya Chakraborty and Prof Samudrala Gourinath as well as all their wonderful students who graciously hosted me in India. I am so grateful for the opportunity to conduct research in such a beautiful setting at Jawaharlal Nehru University. I am particularly thankful for the fruitful discussions.

Prof Lizelle Piater and Dr Gerrit Koorsen for allowing me to use their liquid chromatography system at the University of Johannesburg and to Mr Nicolaas Salomane for kindly assisting me.

I would also like to acknowledge the National Research Foundation South Africa IRG-India-South Africa Cooperative Programme and Blue Skies for providing funding assistance.

I would like to thank my friends and past colleagues who have encouraged and supported me throughout this process. A special thank you to Selisha Naidoo and Tamin Viero – two phenomenal friends who were with me from the day that I embarked on this journey.

To my family –for their generous support, incessant pep talks and for keeping me on my toes! Their love has carried me through the final hurdles.

To the universe, for creating it all.

Dedication

This thesis is dedicated to my mother Hanna, and in memory of my late father, Janusz.

Research Outputs

A. M. Zwolinski and M. E. C. Rey: (Oral presentation) *Structural and Functional Analysis of Tomato curly stunt virus Replication-associated (Rep) Protein.* University of Pretoria Workshop and Mini-symposium on Post-translational Modifications of Proteins and Modification-specific Proteomics, 11th October 2017, Pretoria.

A. M. Zwolinski and M. E. C. Rey: (Oral presentation) *Structural and Functional Analysis of Tomato curly stunt virus Replication-associated (Rep) Protein.* South Africa-India Bilateral Research Co-operation Progress Meeting and Research Visit to the Lab of Prof. S. Chakraborty, Jawaharlal Nehru University, 8th September 2018, Delhi.

A. M. Zwolinski, I. A. Achilonu, and M. E. C. Rey: (Poster Presentation) *Geminivirus Replication-associated Protein: The Path to Solving the Structure.* University of Cape Town Workshop and Mini-symposium on Biophysics & Structural Biology at Synchrotrons, 21st January 2019, Cape Town.

A. M. Zwolinski and M. E. C. Rey: (Oral Presentation) *Structural and Functional Analysis of Tomato curly stunt virus Replication-associated Protein*. South Africa-India Bilateral Research Co-operation Progress Meeting, 21st May 2019, Johannesburg.

A. M. Zwolinski and M. E. C. Rey: (Oral Presentation) *Tomato curly stunt virus – a tail of two variants.* Wits School of Molecular and Cell Biology Virtual Progress Seminar, 18th November 2020.

A. M. Zwolinski and M. E. C. Rey: (Oral Presentation) *Identifying the roles of coding and non-coding regions in the pathogenicity of Tomato curly stunt virus (ToCSV)*. South African Society of Microbiology 2021 Virtual Conference, 5th May 2021.

A. M. Zwolinski and M. E. C. Rey: (Oral Presentation) *Mutational and Structural Analyses of Tomato curly stunt virus Pathogenicity Determinants.* Wits School of Molecular and Cell Biology Virtual Exit Seminar, 12th May 2021.

A. M. Zwolinski, A. Brigden, and M. E. C. Rey: (Publication) *Differences in the 3' intergenic region and the V2 protein of two variants of tomato curly stunt virus play an important role in disease pathology in Nicotiana benthamiana*. Submitted 18th February 2022 to Phytopathology (American Phytopathological Society).

Table of Contents

Contents

List of Figures

331-359). Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved His-TF-RepC and thrombin-cleaved RepC indicated below map. II in green indicates stable, II in red indicates unstable. **B**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced cell lysate (CL), insoluble (INS), and soluble (SOL) fractions extracted from *E. coli* BL21(DE3) pLysS liquid culture transformed with pCold TF-RepC after overexpression at 15 °C for 24 h. Target product band (79.8 kDa) indicated. **C**, Absorbance profile of one-step Ni^{2+} IMAC ÄKTA purification of His-TF-RepC showing peaks corresponding to TWEEN 20 wash, high salt wash, and target elution with buffer containing 500 mM imidazole. **D**, SDS-PAGE analysis of six sequential elution fractions obtained from one-step Ni^{2+} IMAC ÄKTA purification of His-TF-RepC (indicated using arrow) from soluble fraction extracted from E. coli BL21(DE3) pLysS liquid culture transformed with pCold TF-RepC after overexpression at 15 °C for 24 h. M, protein molecular mass marker. Details of nt and predicted aa sequences of recombinantly expressed His-RF-RepC in S3 Fig. 3.1.121

Fig. 3.14 Thrombin cleavage of His-TF-RepC and subsequent purification of RepC protein. **A**, SDS-PAGE analysis of concentrated Ni^{2+} IMAC ÄKTA-purified His-TF-RepC subjected to cleavage with 1 U thrombin at 4 °C for up to 16 h. Cleavage reaction samples were analysed at indicated time intervals. Uncleaved (UC) His-TF-RepC (79.8 kDa), cleaved His-TF (51.4 kDa) tag and RepC (28.4 kDa) bands indicated. **B**, SDS-PAGE analysis of UC and thrombin-cleaved (TC) His-TF-RepC; as well as flow-through (FT) and elution (EL) fractions obtained from one-step Ni^{2+} IMAC ÄKTA purification of RepC after thrombin cleavage for 16 h at 4 °C. UC His-TF-RepC, TC His-TF, and RepC indicated using arrows. M, protein molecular mass marker..... 122

Fig. 3.15 Expression of His-RepC protein. **A**, Map of recombinant His-RepC as expressed using pCold I-RepC. Translation enhancing element (TEE) in yellow, hexahistidine (His6) tag in blue. Factor Xa cleavage site in orange. A 10mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved His-RepC indicated below map. II in red indicates unstable. **B**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced soluble (SOL) fractions extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. Target product band (29.5 kDa) indicated. Non-transformed (NT) soluble fraction included for

protein molecular mass marker. **C**, Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved MBP-RepC_14 and cleaved RepC_14. II in green indicates stable, II in red indicates unstable. **D**, Predicted XtalPred crystallisability classification using random forest classifier (RF) method for cleaved RepC (left) and RepC_14 (right) expressed using pMAL-c2X. 131

List of Tables

List of Symbols and Abbreviations

α	alpha
aa	amino acid
AAV2	adeno-associated virus 2
ACMV	African cassava mosaic virus
ADP	Adenosine diphosphate
AGO	Argonaute
ALCScV	Ageratum leaf curl Sichuan virus
AYVV	Ageratum yellow vein virus
AS	ammonium sulfate
ATP	adenosine triphosphate
ATPγS	adenosine 5'-(gamma-thiotriphosphate)
β	beta
BCTV	beet curly top virus
BLAST	Basic Local Alignment Search Tool
bp	base pair
BPV	bovine papillomavirus
°C	degree Celsius
C6cd	C6/AC6 conserved core domain
CDD	conserved domain database
cDNA	complementary DNA
ChiLCV	chilli leaf curl virus
CLCuBuV	cotton leaf curl Burewala virus
CLCuGeV	cotton leaf curl Gezira virus
CLCuMuV	cotton leaf curl Multan virus
CLE	conserved late element
cm	centimetre
CMD	cassava mosaic disease
CMT3	chromomethylase 3
CR	common region
CP	coat protein
cryo-EM	cryo-electron microscopy

cs	complementary-sense
cspA	cold shock protein A
C-terminal	carboxyl-terminal
C-tT	C-terminal acidic tail
Ct	cycle threshold
CTAB	cetyltrimethylammonium bromide
cTP	chloroplast transit peptide
cv	cultivar
3D	three-dimensional
ΔG	change in Gibbs Free Energy
DaLCV	Datura leaf curl virus
DEAE	diethylaminoethanol
DEPC	diethyl pyrocarbonate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dpi	days post inoculation
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
DTT	dithiothreitol
EACMV	East African cassava mosaic virus
ER	endoplasmic reticulum
FBNYV	faba bean necrotic yellows virus
g	gram
xg	relative centrifugal force
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GFP	green fluorescent protein
GRIK	geminivirus rep-interacting kinase
GRIMP	geminivirus rep-interacting motor protein
GRS motif	geminivirus replication sequence motif
GST tag	glutathione S-transferase tag
GTP	guanosine triphosphate
h (time)	hour (time)
HeuYMV	Hedyotis uncinella yellow mosaic virus

His6 tag	hexahistidine tag
HPV	human papillomavirus
HR	hypersensitive response
Hz	Hertz
ICMV	Indian cassava mosaic virus
ID	identity
II	instability index
IMAC	immobilised metal affinity chromatography
IPTG	Isopropyl ß-D-1-thiogalactopyranoside
IR	intergenic region
IRD	iteron-related domain
kb	kilobase
kcal	kilocalorie
kDa	kilodaltons
LB (medium)	Luria-Bertani (medium)
LIR	long intergenic region
lncRA	long non-coding RNA
M	molar
MCM2	minichromosome maintenance 2
MET1	methyltransferase 1
M-Rep	master Rep
MBP	maltose-binding protein
µg	microgram
miRNA	microRNA
min (time)	minute (time)
µL	microlitre
µM	micromolar
mL	millilitre
mm	millimetre
mM	millimolar
MMT	million metric tons
mol	mole
MP	movement protein

mRNA	messenger RNA
MSV	maize streak virus
mTP	mitochondrial transit peptide
MUSCLE	MUltiple Sequence Comparison by Log- Expectation
MYMIV	mungbean yellow mosaic India virus
NbHDA6	*N. benthamiana* histone deacetylase 6
N-terminal	amino-terminal
ng	nanogram
NLHr	N-terminal long helix region
NLS	nuclear localisation signal
NMR	nuclear magnetic resonance
NSP	nuclear shuttle protein
nt	nucleotide
NTC	no template control
NW	New World
OD	optical density
ORF	open reading frame
Ori	origin of replication
OW	Old World
%	percentage
PaLCuV	papaya leaf curl virus
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PCV2	porcine circovirus 2
PDB	Protein Data Bank
PHYVV	pepper huasteco yellow vein virus
PKC motif	protein kinase C motif
PMSF	phenylmethylsulfonyl fluoride
PRALINE	PRofile Alignment
pRb	retinoblastoma protein
PSIPRED	PSI-blast based secondary structure PREDiction
PTGS	post-transcriptional gene silencing
PVX	potato virus X

qPCR	quantitative PCR
RACE	rapid amplification of cDNA ends
RCA	rolling circle amplification
RDR	RNA-dependent RNA polymerase
REn	replication enhancer protein
Rep	replication-associated protein
RFCp	replication factor C protein
RF method	random forest classifier method
RISC	RNA-induced silencing complex
RITS	RNA-induced transcriptional silencing complex
RM	rich medium
RNA	ribonucleic acid
RPA32	replication protein A32
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription PCR
s (time)	second (time)
SA	South Africa
SACMV	South African cassava mosaic virus
SCE1	small ubiquitin-related modifier-conjugating enzyme
SD	standard deviation
sDBM	single-stranded DNA bridging motif
SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SF3	superfamily 3
SGS3	suppressor of gene silencing 3
siRNA	small interfering RNA
SLCuV	squash leaf curl virus
SOB (medium)	Super Optimal Broth (medium)
SOC (medium)	Super Optimal broth with Catabolite repression (medium)
SP	signal peptide
SPD motif	specificity determinant motif
ss	secondary structure
ssDNA	single-stranded DNA

SUMO	small ubiquitin-related modifier
SV	swollen veins
SV40	simian vacuolating virus 40
T7RNAP	T7 RNA polymerase
TACE	TATA-associated composite element
Tag	large T antigen
TbCSV	tobacco curly shoot virus
TbLCZV	tobacco leaf curl Zimbabwe virus
TF	trigger factor
TGMV	tomato golden mosaic virus
TGS	transcriptional gene silencing
Tm	primer melting temperature
TMH	transmembrane helix
ToCMoV	tomato chlorotic mottle virus
ToCSV	tomato curly stunt virus
ToLCAnV	tomato leaf curl Anjouan virus
ToLCDiV	tomato leaf curl Diana virus
ToLCToV	tomato leaf curl Toliara virus
ToLCGuV	tomato leaf curl Gujarat virus
ToLCGV	tomato leaf curl Ghana virus
ToLCJaV	tomato leaf curl Java virus
ToLCKMV	tomato leaf curl Comoros virus
ToLCMGV	tomato leaf curl Madagascar virus
ToLCMLV	tomato leaf curl Mali virus
ToLCNDV	tomato leaf curl New Delhi virus
ToLCSCV	tomato leaf curl Seychelles virus
ToLCSDV	tomato leaf curl Sudan virus
ToLCUV	tomato leaf curl Uganda virus
ToLDeV	tomato leaf deformation virus
TrAP	transcriptional activator protein
TRV	tobacco rattle virus
TYLCCV	tomato yellow leaf curl China virus
TYLCSV	tomato yellow leaf curl Sardinia virus

TYLCTHV	tomato yellow leaf curl Thailand virus
TYLCV	tomato yellow leaf curl virus
ULR	upward leaf roll
U	unit
USA	United States of America
vs	virion-sense
vsRNA	virus-derived siRNA
WA motif	Walker A motif
WmCSV	watermelon chlorotic stunt virus
WB motif	Walker B motif
WDV	wheat dwarf virus

Abstract

Geminiviruses infect a wide variety of important crop plants causing severe disease resulting in significant yield losses worldwide. Tomato production in the subtropical growing regions of South Africa (SA) is threatened by the emergence of *Tomato curly stunt virus* (ToCSV), a monopartite ssDNA geminivirus belonging to the genus *Begomovirus*. The broad aim of this study was to examine the coding capacity, sequence features, and the role of identified sequence differences on the infectivity in vivo of two variant group isolates (V30 and V22) identified in a previous nationwide survey of tomato (Esterhuizen, 2012). An additional study aim was to explore the structure and investigate the function of ToCSV replication-associated protein (Rep) using in silico and lab-based techniques. Both V30 and V22 DNA-A genomes were predicted to contain three putative complementary-sense (cs) ORFs (C5, C6, C7) and two virion-sense (vs) ORFs (V3 and V4). Sequence alignment of putative ORFs identified in related African mono- and bipartite begomoviruses revealed that the C5/AC5 and C6/AC6 ORFs were the most frequently identified putative ORFs encoding divergent proteins. Two novel putative protein domains were described: a C5/AC5 N-terminal long helix region (NLHr) containing a predicted transmembrane domain, and a C6/AC6 conserved core domain (C6cd). This project was the first to conduct a comprehensive assessment of putative ORFs in African begomoviruses. Additional experiments are necessary to confirm their expression and determine their role in infection.

Previously, an agroinfectious clone of V30 was shown to induce severe disease symptoms, whilst an agroinfectious clone of V22 induced significantly milder disease symptoms in tomato cv Rooi Khaki (Esterhuizen, 2012). Infectivity assays of V30 and V22 were conducted for the first time in *Nicotiana benthamiana* in the current study. The symptom phenotype induced by V30 was characterised by severe swollen veins (SV) and leaf rugosity with no recovery in disease symptoms. V22 induced moderate to severe upward leaf roll (ULR) with a significant recovery in both ULR and SV disease symptoms being observed between 24 and 36 days post inoculation (dpi). Viral load was quantified using qPCR and determined to be significantly higher in V30-infected plants at peak (20 dpi) and late (36 dpi) infection timepoints. Using RT-PCR with strand-specific primers, RNA transcripts derived from both vs and cs strands were

identified in V30-infected plants. The identified transcripts spanned multiple canonical and putative ORFs, putative V2 and C1 5' ORF extension regions, as well as the intergenic region (IR) crossing the origin of replication (Ori). The presence of these transcripts suggests that putative ORFs are transcribed, and the presence of long transcripts (~1.2 kb) may be evidence of bi-directional readthrough transcription to generate dsRNA for downstream RNA silencing pathways, although additional experiments are required to confirm this.

The role of sequence differences in the 3' IR and V2 ORF regions on infectivity of V30 and V22 in *Nicotiana benthamiana* was investigated using V30 partial 3' IR swap mutant chimeras, as well as V30 and V22 canonical V2 ORF swap mutant chimeras. Plants infected with the V30ΔIR-s mutant containing a 78 nt segment of the V22 3' IR developed ULR, a symptom characteristic of V22 infection. A significant increase in viral load at 28 dpi was also observed. Altered symptom phenotypes induced by shorter IR swap mutants suggested that the unique arrangement of a novel *cis*-acting TATA-associated composite element (TACE) situated to the right of the conserved TATA box promoter contributes to the observed phenotype differences. *N. benthamiana* infected with V30ΔV2-s displayed a significant reduction in SV symptom severity at late infection and significantly reduced viral titre was measured at this timepoint compared to that measured in V30-infected plants. The V22ΔV2-s mutant induced significantly increased disease symptom severity which was associated with a significant increase in viral load at peak and late infection. A loss of symptom recovery phenotype was reported with symptom severity remaining high at 36 dpi. Taken together, these findings indicate that the ToCSV 3' IR region is a symptom phenotype determinant in *N. benthamiana* whilst sequence differences present in the canonical V2 ORF are responsible for the modulating viral titre, disease symptom severity, and V22-specific disease recovery phenotype. The precise mechanisms for these observed changes warrant further investigation and additional infectivity assays should be conducted in tomato.

Sequence analysis of the Rep proteins of ToCSV and related African begomoviruses was conducted and the presence of key HUH endonuclease motifs (motif I, II, III), and ATPase/superfamily 3 (SF3) helicase motifs (Walker A, Walker B, B' motif, C motif) were confirmed. Homology modelling of ToCSV Rep showed that the N-terminal endonuclease domain (monomer) was predicted to contain an antiparallel β-sheet

(five β-strands), whilst the C-terminal ATPase/helicase domain (homohexamer) contained a parallel β-sheet (four β-strands). The arrangement of important functional motifs was similar to that present in the solved structures of other DNA virus Rep proteins. Agroinoculation of *N. benthamiana* with V30 C1 deletion mutants indicated that deletion of the 14 C-terminal residues of Rep (aa 346 – 359) resulted in significantly reduced viral load and greatly delayed symptom development. Removal of the 22 C-terminal residues of Rep (aa 338 – 359) abolished infectivity entirely. From these results it was concluded that the 14 C-terminal residues (predicted to be disordered) are not essential for Rep enzymatic activity which is a prerequisite for successful viral replication in vivo. The ToCSV RepC protein (aa 122 – 359) was successfully expressed in *Escherichia coli* as a recombinant fusion. Additional optimisation of expression and purification is required to obtain sufficient purified protein for future structural studies. Miniaturised expression trials of RepC truncated by 14 C-terminal residues appeared to yield higher quantities of protein compared to full length RepC as evidenced by SDS-PAGE. Targeted removal of disordered C-terminal residues is therefore suggested as a practical approach to improve yield of recombinantly expressed RepC whilst retaining the enzymatic activity of the domain of interest.

The present study contributed to increasing the knowledge of putative begomovirus ORFs as well as providing insight into the possible functions of selected sequence elements of interest present in the ToCSV genome. It is anticipated that the broad scope of this study will generate several intriguing avenues for future research of ToCSV and related African begomoviruses which remain largely understudied. It is hoped that this research will contribute to improving the understanding of these pathogens with the aim of developing measures aimed at protecting the expanding agriculture industry in Southern Africa.

Chapter 1: Literature Review

1.1 Tomato

1.1.1 The crop

Tomato (*Solanum lycopersicum* L.) is a versatile horticultural and agricultural food crop belonging to the dicotyledonous plant family Solanaceae. The tomato plant is grown exclusively for its fruit which is considered a culinary vegetable, ranking second in economic importance as a solanaceous crop after potato (Simonne et al., 2011). The centre of origin is believed to be South America (Bolivia and Peru). The Spanish Conquest of the Americas led to the introduction of the tomato into Europe from where it has since spread worldwide and is now grown in nearly every single country on Earth (Caicedo and Peralta, 2013; Nonnecke, 1989). The tomato can be grown in a wide range of soil types, in the field as well as in the greenhouse on a commercial scale as well as for small-scale home growing (Jones, 2007). There are over 7,500 varieties of tomato and the fruit is rich in vitamins, minerals and carotenoids (Raiola et al., 2014). The tomato may be consumed raw or cooked and may be processed into sauces, preserves, juice and concentrate (Blancard et al., 2012).

1.1.2 Production

The global tomato production quantity was recorded as 186.82 MMT for 2020. Asia was responsible for the production of 62.6 % (116.99 MMT), with China being the top producer accounting for just under 35 % of the worldwide total production (64.86 MMT). The Americas produced 13.1 % (24.44 MMT), Europe produced 12.2 % (22.81 MMT) and 0.2% (0.34 MMT) was produced in Oceania (FAOSTAT, 2022). Africa was responsible for 11.9% (22.22 MMT) of the 2020 production quantity, with Nigeria being the top producer in sub-Saharan Africa (3.69 MMT). South Africa (SA) produced 0.58 MMT of tomatoes in 2020 (FAOSTAT, 2022). In SA, the commercial industry is composed of almost 700 producers and as of 2011, supplied 95% of the total annual fruit output and employed over 22,000 people (DAFF, 2019). In SA, tomato is grown in both summer and winter with 75% of commercially cultivated tomato area being in the Limpopo Province (DAFF, 2019). With the exclusion of potato, the contribution of tomato production in SA constituted ~17 % of vegetable production gross value

reported in 2018 (DAFF, 2019). As tomato is largely harvested by hand, the industry provides employment opportunities for low-skilled agricultural labourers (Heuvelink and Dorais, 2005). Tomato is also an important subsistence crop in sub-Saharan Africa. Many cultivars are high yielding and farmers are able to consume a portion of the fruit they produce as well as sell a portion of their yield at local markets (Odame et al., 2009).

1.2 Geminiviruses infecting crop plants

1.2.1 *Geminiviridae*: a brief description

Comprising of over 500 recognised species, *Geminiviridae* is a family of ssDNA viruses which infect a wide range of plant hosts worldwide (Fiallo-Olivé et al., 2021; Zerbini et al., 2017). This family is so named because of the characteristic viral capsid shape which comprises of two incomplete icosahedra fused to form a twinned (geminate) particle with a size of ~20 x 35 nm which encloses a single circular ssDNA (Hesketh et al., 2018; Zerbini et al., 2017). The complex capsid subunit conformations as well as the interactions between these subunits and the viral genome have been described using structures of geminivirus particles solved using cryo-electron microscopy (cryo-EM) (Hesketh et al., 2018; Hipp et al., 2017). Several hemipteran insects which are known pests of a wide range of important crop plants act as vectors which facilitate the spread of geminiviruses (Perilla-Henao and Casteel, 2016). Until recently, nine distinct genera were recognised: *Becurtovirus*, *Begomovirus*, *Capulavirus*, *Curtovirus*, *Eragrovirus*, *Grablovirus*, *Mastrevirus*, Topocuvirus and *Turncurtovirus* (Zerbini et al., 2017). These genera are differentiated by their nt sequence identity, genome organisation, plant host, and insect vector (Fiallo-Olivé et al., 2021). Recent advances in DNA sequencing and development of high-throughput methods in conjunction with environmental sampling have led to the discovery of a number of unique geminiviruses (Bernardo et al., 2013; Claverie et al., 2018; Roumagnac et al., 2021). Sequence analysis of these divergent geminiviruses identified in different monocotyledonous and dicotyledonous plants resulted in the creation of five novel genera: *Citlodavirus*, *Maldovirus*, *Mulcrilevirus*, *Opunvirus*, and *Topilevirus* (Roumagnac et al., 2021).

It is worth mentioning that the classification of geminiviruses will undoubtedly undergo revision as additional sequence information is made available and currently four highly divergent species remain unassigned to the 14 recognised genera (Roumagnac et al., 2021). The characteristics of the 14 established genera as defined by Fiallo-Olivé et al. (2021), Roumagnac et al. (2021), and Zerbini et al. (2017) are summarised in Table 1.1.

Table 1.1 Summary of recognised genera belonging to the ssDNA virus family Geminiviridae

Genus	Species	Genome size	Plant host	Insect vector
Becurtovirus	3	DNA-A: ~2.8 kb	dicot	leafhopper
Begomovirus	445	DNA-A: ~2.6 kb (monopartite) ~2.7 kb (bipartite) DNA-B: ~2.6 kb	dicot	whitefly
Capulavirus	4	DNA-A: ~2.6 – 2.8 kb	dicot	aphid
Citlodavirus	4	DNA-A: ~3.6 – 3.7 kb	dicot	unknown, possibly whitefly
Curtovirus	3	DNA-A: ~3.0 kb	dicot	leafhopper
Eragrovirus	1	DNA-A: 2.7 kb	monocot	unknown
Grablovirus	3	DNA-A: ~3.2 kb	dicot	treehopper
Maldovirus	3	DNA-A: ~2.7 – 2.9 kb	dicot and monocot	unknown
Mastrevirus	45	DNA-A: ~2.6 – 2.8 kb	dicot and monocot	leafhopper
Mulcrilevirus	2	DNA-A: ~3.0 kb	dicot	leafhopper
Opunvirus	1	DNA-A: 2.9 kb	dicot	unknown, possibly cochineal insect
Topilevirus	2	DNA-A: ~2.6 – 2.8 kb	dicot	unknown, possibly treehopper
Topocuvirus	1	DNA-A: 2.8 kb	dicot	treehopper
Turncurtovirus	3	DNA-A: ~2.9 kb	dicot	leafhopper

Remarkably, SA has proven to be somewhat of a hotspot for geminivirus diversity. Environmental sampling of non-crop plants native to Southern Africa has led to the identification of highly divergent geminiviruses: *Exomis microphylla latent virus* (genus *Becurtovirus*), *Euphorbia caput-medusae latent virus* (genus *Capulavirus*), *Eragrostis curvula streak virus* (genus *Eragrovirus*), *Limeum africanum associated virus* (genus unassigned), and *Polygala garcinii associated virus* (genus unassigned) (Bernardo et

al., 2013; Claverie et al., 2018; Varsani et al., 2009). Five distinct species belonging to the genus *Mastrevirus* have also been identified in SA infecting French bean (*Phaseolus vulgaris* L.), sugar cane (genus *Saccharum*), maize (*Zea mays* L.) and related grasses (Fiallo-Olivé et al., 2021; Lawry et al., 2009; Lazarowitz, 1988; Liu et al., 1997; Varsani et al., 2008). The presence of highly divergent geminiviruses has significant implications for their continued evolution in both wild and crop plant hosts.

The majority of recognised geminivirus species belong to the genus *Begomovirus* which infects dicotyledonous plants and is well represented in both the New World (NW: the Americas) and the Old World (OW: Afro-Eurasia). Begomoviruses may have a monopartite genome (DNA-A only), or a bipartite genome composed of DNA-A and DNA-B (Fiallo-Olivé et al., 2021). The DNA-A component contains five recognised ORFs that are found in all begomoviruses. The virion-sense (vs) ORF V1 (AV1 in bipartite begomoviruses) encodes the coat protein (CP). The four complementary-sense (cs) ORFs and their encoded proteins include: C1 (replication-associated protein – Rep), C2 (transcriptional activator protein – TrAP), C3 (replication enhancer protein – REn), and C4 (C4 protein). These ORFs are designated AC1, AC2, AC3 and AC4 in bipartite begomoviruses (Fiallo-Olivé et al., 2021). An additional vs ORF, V2/AV2 is found in OW and Australian begomoviruses only (Chowda-Reddy et al., 2008; Nawaz-ul-Rehman and Fauquet, 2009). The DNA-B component of bipartite begomoviruses contains two recognised ORFs: BV1 encoding the nuclear shuttle protein (NSP), and BC1 encoding the movement protein (MP) (Fondong, 2013).

Geminivirus proteins have been shown to be highly multifunctional and are known to interact with nucleic acids, host and other viral proteins to facilitate viral DNA replication, movement in the host and vector transmission (Devendran et al., 2022). The CP is sole structural protein which forms the geminivirus capsid. In monopartite geminiviruses the CP also facilitates the movement of viral DNA from the nucleus to the cytoplasm and contains DNA binding and nuclear localisation motifs in the N-terminal region (Fondong, 2013; Sharma and Ikegami, 2009; Unseld et al., 2001). The CP of monopartite begomoviruses is thus a functional homologue of the BV1 protein of bipartite begomoviruses (Fondong, 2013; Rojas et al., 2001). The CP also facilitates inter-cellular movement of monopartite begomoviruses (Hehnle et al., 2004). Vector transmission is mediated by the CP and vector specificity has been shown to be determined by specific sequences in the CP (Devendran et al., 2022). The V2 protein

has several functions related to pathogenicity and RNA silencing suppression and is described in greater detail in Section 1.3.2. The geminivirus Rep protein is a multifunctional protein which is essential for successful viral replication (Fondong, 2013; Hanley-Bowdoin et al., 2013). The structure and function of ToCSV Rep is discussed further in Section 1.4. The TrAP of tomato yellow leaf curl virus (TYLCV) can bind both ss and dsDNA in a sequence non-specific manner (Noris et al., 1996). In the monopartite and bipartite begomoviruses, TrAP has been shown to function as a transcriptional activator necessary for the expression of vs ORFs (Devendran et al., 2022; Fondong, 2013; Hanley-Bowdoin et al., 2013). Geminivirus TrAP has also been implicated in the suppression of RNA silencing (Castillo-Gonzáles et al., 2015; Trinks et al., 2005; Gover et al., 2014; Kon et al., 2007; Peretz et al., 2011; Sundaresan et al., 2020). The REn protein is known to interact with Rep and enhance viral replication (Fondong, 2013). It has been revealed that the REn of the monopartite tomato yellow leaf curl Sardinia virus (TYLCSV), in addition to interacting with Rep, is able to undergo interaction with the proliferating cell nuclear antigen (PCNA). The PCNA is a critical host protein component of the plant cell replication machinery which begomoviruses use to replicate in the host (Castillo et al., 2003). The C4 protein of monopartite begomoviruses acts as a symptom determinant (Teng et al., 2010) and it has also been linked to suppression of RNA silencing and viral DNA movement (Melgarejo et al., 2013; Fondong, 2019). In African cassava mosaic virus (ACMV) infection, AC4 is able to bind microRNAs (miRNAs) thus suppressing the RNA silencing pathway mediated by the RNA-induced silencing complex (RISC) (Chellappan et al., 2005; Liu et al., 2017). The two proteins encoded by DNA-B of bipartite begomoviruses are necessary for inter- and intracellular viral movement. NSP binds viral DNA and shuttles it from the nucleus into the cytoplasm. By modifying the plasmodesmatal size exculsion limit, the BC1 protein facilitates movement of viral DNA between adjacent cells (Rojas et al., 1998). NSP is believed to be critical in allowing the movement of viral DNA from the cell nucleus through the nuclear envelope and into the cytoplasm (Fondong, 2013; Hehnle et al., 2004). NSP binds both ss and dsDNA in a sequence non-specific fashion (Hehnle et al., 2004). The MP enables intercellular as well as long distance movement of viral DNA by binding to the NSP-DNA complex (Fondong, 2013). MP is hypothesized to alter the plasmodesmata to allow the DNA-protein complex to move between cells (Gafni and Epel, 2002).

The non-coding region between the cs and vs ORFs of DNA-A which contains the origin of replication (Ori) is termed the intergenic region (IR). The nt sequence between BV1 and BC1 in DNA-B is referred to as the long intergenic region (LIR) whilst a conserved sequence present in both the DNA-A IR and DNA-B LIR of bipartite begomoviruses is known as the common region (CR) (Zerbini et al., 2017). Begomoviruses have been shown to associate with subviral circular ssDNA molecules which serve to modify disease symptom severity and alter viral titre in the host (Briddon et al., 2012; Devendran et al., 2022). Three classes of these subviral agents have been identified: alphasatellites, betasatellites and deltasatellites. The alphasatellites (~1.2 – 1.4 kb) are not considered to be strict satellite DNAs since they encode a Rep protein and are thus able to replicate in the absence of the helper virus (Briddon et al., 2018; Lozano et al., 2016). The helper virus is instead required for movement within the infected host and insect vector transmission (Mansoor et al., 1999; Nawaz-ul-Rehman and Fauquet, 2009). The Rep of alphasatllites has been shown to suppress RNA silencing (Abbas et al., 2019). Betasatellites (~1.3 kb) and deltasatellites (~700 nt) require their helper virus for replication (Nawaz-ul-Rehman et al., 2021). Betasatellites contain a cs ORF (βC1) encoding a protein that acts as a pathogenicity determinant and suppressor of host immunity pathways (Gnanasekaran et al., 2019). The recently-identified vs ORF (βV1) encodes a protein which acts as a pathogenicity determinant (Hu et al., 2020). Deltasatellites (~700 nt) are apparently non-coding but have been shown to modulate the viral DNA accumulation of their helper viruses (Ferro et al., 2021; Lozano et al., 2016).

1.2.2 Begomoviruses: a pertinent threat

Begomoviruses are important pathogens of cultivated tomato. The TYLCV-related complex contains over 35 closely related tomato begomoviruses (Zerbini et al., 2017). Two strains of TYLCV (mild and Israel strains) infect tomato in tropical and sub-tropical regions worldwide and this virus is one of the most devastating pathogens of tomato (Mabvakure et al., 2016). TYLCV can induce up to 100% yield losses in some tomato cultivars depending on the growth stage of the plant (Lapidot et al., 2001; Pan et al., 2012). It is important to note that the begomoviruses in the TYLCV-related complex are prone to recombination events in mixed infections which gives rise to new variants and strains (Moriones et al., 2007; Padidam et al., 1999). Mixed infection of distinct

begomoviruses contributes significantly to increasing their variability as well as being a major driver of their evolution. Morilla et al. (2004) presented evidence of the DNA-A components of two distinct tomato-infecting begomoviruses being colocalised within the nuclei of infected phloem parenchyma and companion cells when in mixed infection. Begomoviruses replicate in the cell nucleus and such colocalization will inevitably result in recombination events as well as pseudo-recombination in bipartite viruses where the reassortment of genomic components occurs (Esmaeili et al., 2015). Pseudo-recombination resulting in increased symptom severity has been observed in bipartite tomato-infecting begomoviruses (Chakraborty et al., 2008; Kanakala et al., 2013). Heteroencapsidation, where the DNA component of one virus is packaged into the CP of another, is another demonstrated outcome of mixed begomovirus infection and has consequences for vector transmission (Kanakala et al., 2013).

Viral synergism, recombination, and pseudo-recombination events among several species of almost exclusively bipartite cassava-infecting begomoviruses have been widely reported (Jacobson et al., 2018; Pita et al., 2001; Rey and Vanderschuren, 2017; Vanitharani et al., 2004; Zinga et al., 2013). These events are the major driving force behind the development of the cassava mosaic disease (CMD) epidemic in sub-Saharan Africa which has caused catastrophic losses in tuber yield (Jacobson et al., 2018; Rey and Vanderschuren, 2017). Spread of begomoviruses that cause CMD is facilitated not only by the spread of the whitefly vector, but also by the uncontrolled movement of infected material through vegetative propagation (Legg et al., 2014). Cassava infected with CMD begomoviruses can incur up to 40% yield losses and the total annual yield losses in Africa due to CMD is estimated to be over 45 MMT (Thresh et al., 1997). A distinct bipartite cassava infecting *Begomovirus* species, *South African cassava mosaic virus* (SACMV), was identified in cassava displaying symptoms of CMD in Swaziland and the subtropical eastern regions of SA (Berrie et al., 1998). Although SACMV was shown to be closely related to viruses in the TYLCV-related complex based on CP aa sequence and CR nt sequence comparisons, it was serologically more closely related to *East African cassava mosaic virus* (EACMV) (Berrie et al., 1998). Due to the unique sequences contained within DNA-A, it has been suggested that SACMV emerged as a result of recombination events between at least two different begomoviruses including an unknown begomovirus species (Berrie et al., 2001).

The evolutionary relationships of begomovirus populations present in important cultivated crops such as tomato, tobacco and cassava in sub-Saharan Africa and Indian Ocean islands in the geographical proximity to the continent have been previously investigated (Lefeuvre et al., 2007a; Leke et al., 2015; Mivedor et al., 2017; Rey et al., 2012). Several distinct African begomovirus species are recognised but additional surveys are necessary to characterise the rapid evolution of these viruses, especially in non-cultivated weed and indigenous plant hosts (Rey et al., 2012).

1.2.3 Tomato curly stunt virus (ToCSV)

Tomato curly stunt virus (ToCSV) is a distinct tomato infecting *Begomovirus* species and was first identified in 1997 in commercial tomato growing in the Onderberg region of Mpumalanga. By 2003, ToCSV had spread to cultivated tomato in Mozambique as well as the subtropical growing regions of SA (KwaZulu-Natal and Limpopo Provinces) (Pietersen et al., 2008). ToCSV induces similar symptoms to TYLCV in tomato, including yellowing of the young upper leaves, curling and cupping of leaflet margins, severe stunting and dwarfing of overall plant growth as well as decreased fruit set (Pietersen et al., 2000, 2008). Like TYLCV, this begomovirus can induce up to 100% yield losses in susceptible cultivars (Pietersen and Smith, 2002). A study of several tomato cultivars which are known to show resistance to TYLCV were found to also show resistance to ToCSV but none of the cultivars were immune to ToCSV infection (Pietersen and Smith, 2002).

ToCSV was determined to be closely related to the TYLCV-related complex based on the comparison of CP gene sequences, with the TYLCV-Israel strain sharing 76.2% sequence identity (Pietersen et al., 2000). Phylogenetic analysis revealed that the ToCSV genome shares ~84% sequence identity with the DNA-A of tobacco leaf curl Zimbabwe virus (TbLCZV). Interestingly, in silico recombination analysis revealed the presence of two recombinant fragments in the ToCSV genome which were likely sourced from closely related relatives of SACMV and EACMV. This suggests that southern African cassava-infecting begomoviruses and ToCSV may infect a common host, thus allowing recombination to occur (Pietersen et al., 2008). Since ToCSV is closely related to southern African begomoviruses infecting solanaceous plant hosts,

it has been suggested that ToCSV emerged locally as a distinct virus indigenous to southern Africa (Pietersen et al., 2008).

The ssDNA genome component (Fig. 1.1) of ToCSV (AF261885) is 2,766 nt in length and contains an IR (259 nt) between ORFs C1 and V2. The IR contains the Ori situated on the conserved stem-loop nonanucleotide sequence 5'-TAATATTAC-3' as well as Rep-binding iteron sequence sites (Pietersen et al., 2008). ToCSV is a monopartite begomovirus, as no DNA-B component was located using universal DNA-B PCR primers. In addition, no satellite DNA molecules were previously found to be associated with ToCSV (Pietersen et al., 2008). The DNA-A of ToCSV contains six accepted ORFs (Esterhuizen, 2012). Two are located on the vs strand: V1 encoding the CP (258 aa) and V2 encoding the V2 protein (113 aa). Four ORFs are located on the cs strand: C1 encoding the Rep protein (359 aa), C2 encoding the TrAP protein (135 aa), C3 encoding the REn protein (134 aa) and C4 encoding the C4 protein (85 aa).

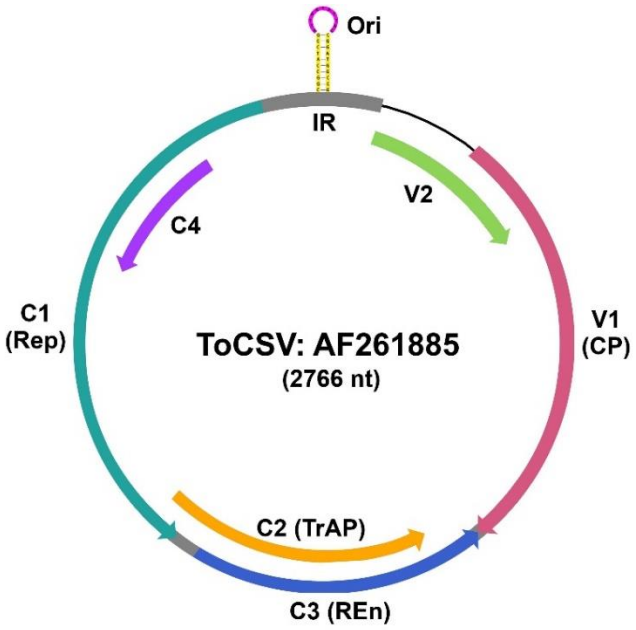


Fig. 1.1 Genome organisation of ToCSV exemplar isolate (AF261885) DNA-A (2766 nt) containing six open reading frames: C1 (Replication-associated protein – Rep), C2 (Transcriptional activator protein – TrAP), C3 (Replication enhancer protein – REn), C4, V1 (Coat protein – CP), V2. Intergenic region (IR) with stem-loop containing the origin of replication (Ori) shown.

A study of ToCSV in the tomato-producing Chókwè District of Mozambique identified two distinct strain groups ToCSV-Chok I and ToCSV Chok-II (Sande et al., 2021). Evidence of recombination in the C1 gene was presented and it was concluded that

speciation was taking place in Mozambique which has implications for tomato producers in this region (Sande et al., 2021). The first serious survey of ToCSV in SA by Esterhuizen (2012) resulted in the discovery of two distinct ToCSV variant groups (ToCSV-I and ToCSV-II). The variant groups were differentiated by disease severity with ToCSV-I isolates causing significantly more severe disease in tomato than ToCSV-II isolates. A recombinant fragment was present in isolates belonging to ToCSV-II spanning nt 109 – 322 which corresponds to the 3' IR and the 5' V2 ORF (Esterhuizen, 2012). The minor parent in this event was determined to be an isolate of tomato leaf curl Uganda virus (ToLCUV) which was surprising since ToLCUV has not been identified in SA (Esterhuizen, 2012).

1.3 Intergenic region (IR) and V2 coding region.

1.3.1 The intergenic region (IR)

The IR of geminiviruses, which is believed to be non-coding in nature, lies between the vs and cs strand coding regions of begomoviruses. It contains a stem-loop element with a nonanucleotide sequence (5'-TAATATTAC-3') which is conserved amongst most geminivirus genera. The nonanucleotide sequence contains the Ori where cleavage by the Rep protein initiates the replication cycle (Akbar et al., 2012; Piedra-Aguilera et al., 2019). Several IR promoter regions that drive the expression of downstream viral genes, as well as IR *cis*-acting elements involved in regulating their transcription have been studied in a wide range of geminiviruses. These elements are essential for efficient bidirectional transcription of the vs and cs ORFs. The mechanisms involved are still being elucidated as the transcriptional activities of geminiviruses are facilitated by complex interactions of several viral and host proteins. (Borah et al., 2016; Kumar and Shivaprasad, 2020). High genomic variability necessitates additional mapping of transcripts, promoters, and regulatory elements in geminiviruses (Borah et al., 2016). It is worth mentioning that geminiviral promoters are not only located in the IR. Several sequences with promoter activity have been identified within coding regions suggesting that other promoter/regulatory elements present outside of the IR may have important function and should be further investigated (Borah et al., 2016; Gong et al., 2021). A study of ACMV by Zhan et al. (1991) revealed that a sequence upstream of the AC2 ORF was able to function as a promoter of moderate strength. A strong unidirectional promoter termed AV3, was

identified in the coding region of the monopartite Ageratum yellow vein virus (AYVV). The promoter was experimentally shown to have activity in both prokaryotic and eukaryotic systems. The core sequence was mapped to a region between nt 762 and 869, which overlaps with part of the C-terminal half of the V1 ORF (Wang et al., 2013; Wang et al. 2014a).

The IR left side promoters driving the expression of early cs ORFs are known to be stronger than the promoters present in the right side of the IR which are responsible for driving late vs ORF expression (Borah et al., 2016; Zhan et al., 1991). Despite the nt sequence of the IR being highly variable amongst geminiviruses, several conserved sequence motifs have been identified. Several short repeat sequences are known to enhance transcription. Direct repeat sequences which may have small internal sequence palindromes have been suggested to exert greater influence than inverted repeat sequence elements (Velten et al., 2005). Both the IR sequences to the left and right of the nonanucleotide each contain a TATA box which functions as a known core promoter and is indispensable to geminivirus replication (Eagle and Hanley-Bowdoin, 1997; Khan et al., 2015). Numerous other important *cis*-acting elements have been mapped and characterised for selected begomoviruses and these include among others: GC box, G box motif (CACGTG), CA motif (CCAAAA), AG motif, and CAAT box (Ashraf et al., 2014; Borah et al., 2016; Khan et al., 2015; Shivaprasad et al., 2005). The left side of the IR contains a series of repeat sequence elements (iterons) which serve as important binding sites for the Rep protein. Iteron sequences vary among geminiviruses but are crucial for successful replication (Xu et al., 2019). The iterons of ToCSV are described in greater detail in Chapter 3.

The right side of the IR contains the late elements necessary for transcription of the vs ORFs. The expression of these late ORFs is dependent upon the activity of the TrAP (Sunter and Bisaro, 1997). A stretch of 15 aa present in the C-terminal acidic region of tomato golden mosaic virus (TGMV) TrAP are required for transcriptional activation (Hartitz et al., 1999). The conserved late element (CLE) is a TrAP-responsive *cis*-acting element present in the right side of the geminivirus IR and has a consensus sequence GTGGTCCC (Borah et al., 2016; Ruiz-Medrano et al., 1999; Shivaprasad et al., 2005; Velten et al., 2005). The CLE has been implicated in the transcriptional activation of the late genes of several, but not all, geminiviruses (Borah et al., 2016; Kumar and Shivaprasad, 2020). Rather than binding to the CLE directly,

it is proposed that TrAP interacts with and manipulates target host factors that assemble at the CLE due to sequence-specific binding. This is supported by experimental evidence indicating that TrAP non-specifically binds ssDNA and interacts minimally with dsDNA also without sequence-specific affinity (Hartitz et al., 1999; Sung and Coutts, 1996).

The number of CLE present in the IR varies, with some geminiviruses having two or three such elements in the right side of the IR (Bañuelos-Hernández et al., 2012). The targeted removal of CLE has been shown to result in the reduced activity of the late promoter in some geminiviruses (Hur et al., 2008; Kumar and Shivaprasad, 2020). Increasing the number of CLE has shown to result in either increased or unaltered promoter activity (Cantú-Iris et al., 2019; Hur et al., 2008). A novel *cis*-acting element termed the TATA-associated composite element (TACE) has been recently described and is present in NW and OW begomoviruses (Cantú-Iris et al., 2019). This element is discussed further in Chapter 2.

1.3.2 V2 ORF

The V2/AV2 ORF is only found in OW geminiviruses (Padidam et al., 1996). The protein encoded by this ORF is known to play a key role in virus pathogenicity and exhibits a multifunctional nature that contributes to virus-host interactions. The V2 of TYLCV has been demonstrated to bind genomic DNA (Moshe et al., 2015) Early studies of monopartite geminiviruses provided evidence that an intact V2 protein is required for successful systemic infection, implicating it in facilitating movement within the host (Padidam et al., 1996; Rigden et al., 1993; Stanley et al., 1992; Wartig et al. 1997). On the contrary, Hak et al. (2015) have shown that TYLCV-Is (Israel isolate) V2-defective mutants are still able to induce systemic infection suggesting that it is not required for movement. Subcellular localisation studies using fluorescence microscopy have revealed that V2/AV2 is primarily localised to the cell cytoplasm, often forming distinct protein aggregates (Chowda-Reddy et al., 2008; Glick et al., 2008; Hak et al., 2015; Sharma et al., 2010; Zhang et al., 2012; Zhao et al., 2018; Zrachya et al., 2007). Localisation of V2/AV2 to the endoplasmic reticulum (ER), nucleus and nuclear periphery has also been observed (Hak et al., 2015; Kumar, 2019; Luna et al., 2017, 2020; Roshan et al., 2017). The V2 protein of TYLCV is known to

interact with itself in the Cajal bodies of *Nicotiana benthamiana* cell nuclei, and Wang et al. (2020) found that this self-interaction is dependent upon a leucine at aa position 76. The putative self-interaction motif of TYLCV V2 spanning aa 61 to 78 was implicated in facilitating self-interaction with a serine residue at aa position 71 being indispensable for self-interaction (Zhao et al., 2018). The authors determined that mutation of a serine to alanine at aa position 71 terminated the ability of V2 to interact with itself. It is intriguing to note that an alanine exists at aa position 71 in ToCSV V2 which is unusual given the findings by Zhao et al. (2018). Since self-interaction is necessary for V2 function, this implies that the alanine at position 71 of wild type ToCSV V2 does not impair its functionality as is the case in TYLCV V2. This suggests that additional residues may be necessary for self-interaction or that serine at aa position 71 may not be a strict requirement for self-interaction of ToCSV and possibly other monopartite begomovirus V2 proteins.

An arginine at aa position 43 has been shown to be required for tobacco curly shoot virus (TbCSV) V2 self-interaction (Li et al., 2021a). Reports of varying subcellular localisations is not surprising given that in addition to self-interaction, V2/AV2 has been found to interact with several other geminiviral proteins. Interactions between V1 and V2 have been implicated in movement of genomic DNA out of the cell nucleus as well as to adjacent cells (Poornima Priyadarshini et al., 2011; Zhao et al., 2020). The C4 protein of monopartite begomoviruses has been implicated in coordinating with V2 and V1 to modify the size-exclusion limit of plasmodesmata to allow for intercellular movement (Kumar, 2019). V2/AV2 is gaining increasing attention for its interactions with various host proteins to modulate disease and interfere with plant antiviral pathways. Many such interactions with host proteins known to function in disease-related pathways have been demonstrated using the yeast two hybrid system (Roshan et al., 2017).

Several geminivirus V2/AV2 proteins have been demonstrated to function as pathogenicity determinants necessary for inducing severe disease (Amin et al., 2011; Chowda-Reddy et al., 2008; Luna et al., 2020; Mubin et al., 2010; Roshan et al., 2018; Sharma and Ikegami, 2010; Sharma et al., 2010; Zhang et al., 2012). Potato virus X (PVX)-vector based expression of several geminivirus V2/AV2 ORFs in *N. benthamiana* has demonstrated that V2/AV2 is able to induce severe virus-like symptoms such as leaf curling and chlorosis. In addition, it has also been shown to

function as an avirulence factor that promotes the development of a hypersensitive response (HR) (Luna et al., 2020; Mubin et al., 2010; Saeed et al., 2018; Sharma and Ikegami, 2010). Deletion mutants of the tomato leaf curl Java virus (ToLCJaV) V2 protein expressed using a PVX-vector displayed altered abilities to induce necrosis and oxidative burst in *N. benthamiana* which are typical HR-like responses. HR induction was lost upon deletion of 58 C-terminal (but not 58 N-terminal) V2 aa which suggests that the C-terminal region of V2 is responsible for elicitation of an HR-like response (Sharma and Ikegami, 2010). Although PVX-vector expression shows that V2 can induce HR, it has also been reported that the TYLCV V2 has the ability to suppress the activity of CYP1, a host papain-like cysteine protease which functions in HR pathways to defend host tissue against infection (Bar-ziv et al., 2012; Bar-ziv et al. 2015). The result is that the activity of V2 suppresses this host defence mechanism.

V2/AV2 has demonstrated the ability to function as a potent suppressor of two classes of RNA silencing by several modes of action. Post-transcriptional gene silencing (PTGS) is triggered by the presence of dsRNA intermediates which are cleaved by the host Dicer-like ribonucleases into small interfering RNAs (siRNAs) 20-24 nt in length. These siRNAs associate with Argonaute (AGO) proteins and other host factors to generate RISC which targets virus-derived mRNAs in a sequence-specific manner which results in their degradation to prevent their translation (Ashfaq, 2020; Bisaro, 2006; Hoffer, 2011). In transcriptional gene silencing (TGS) the RNA-induced transcriptional silencing complex (RITS) containing siRNAs scavenges dsDNA containing complementary sequences (in this case circular dsDNA replicative form of geminivirus genome) and inhibits their transcription (Bisaro, 2006; Rishishwar and Dasgupta, 2018) Methylation of viral DNA is an important TGS defence mechanism employed by the host to counter viral infection and geminivirus V2/AV2 has been shown to function as a suppressor of TGS (Mubin et al., 2019; Wang et al., 2014b, 2019). Evidence indicates that the TGS suppression mechanism of TYLCV V2 in *N. benthamiana* involves competitive protein binding. V2 interacts with *N. benthamiana* histone deacetylase 6 (NbHDA6) to impede its interaction with methyltransferase 1 (MET1). The interaction of these two host proteins is necessary for viral DNA methylation and the competitive binding of V2 with NbHDA6 (but not MET1) resulted in suppression of this methylation pathway (Wang et al., 2018). The V2 proteins of both TYLCV and cotton leaf curl Multan virus (CLCuMuV) are known to interact with

AGO4 in the Cajal bodies of host cell nuclei, with AGO4 being a key host component in RNA-directed DNA methylation (Wang et al., 2019; Wang et al. 2020). This interaction prevents AGO4 from binding to viral DNA. The presence of a single leucine residue at aa position 76 is necessary for the interaction of CLCuMuV V2 and AGO4. Interestingly, mutation of leucine 76 does not disrupt the interaction between TYLCV V2 and AGO4, although this residue is required for TYCLV V2 self-interaction in the Cajal body (Wang et al., 2019; Wang et al., 2020). This strongly suggests that sequence variations in geminivirus V2 proteins may significantly contribute to altering their apparently diverse mechanisms of action in TGS suppression and that much remains to be discovered in this regard. Geminivirus V2 proteins have also been shown to suppress PTGS at various stages. Host RNA-dependent RNA polymerase 6 (RDR6) coordinates with the dsRNA-binding protein suppressor of gene silencing 3 (SGS3) for the generation of short interfering RNA (siRNA) necessary for RNA silencing. TYLCV V2 has been found to interact and bind with tomato SGS3, as well as outcompete SGS3 by binding dsRNA thus preventing downstream siRNA generation (Glick et al., 2008; Fukunaga and Doudna, 2009). The V2 of tomato yellow leaf curl China virus (TYLCCV) was not able to bind tomato SGS3 in a yeast two hybrid system but was instead shown to directly bind specific classes of siRNAs suggesting that siRNA sequestration by V2 is another mechanism used by geminiviruses to suppress PTGS (Zhang et al., 2012). Two putative hydrophobic domains in the N-terminal region of beet curly top virus (BCTV) V2, H1 and H2, were shown to be necessary for RNA silencing suppression and systemic movement (Luna et al., 2020).

Begomovirus V2/AV2 proteins are known to have two key sequence motifs necessary for specific functions. The putative protein kinase C (PKC) phosphorylation motif has a consensus sequence of Thr-x-Arg, where x can be any aa residue. An intact PKC motif is essential for the ability of EACMV AV2 to function as a pathogenicity determinant (Chowda-Reddy et al., 2008). A mutant papaya leaf curl virus (PaLCuV) V2 protein with a partial deletion lacking this motif was found to have impaired HR elicitation in PVX-vector based expression in *N. benthamiana* but specific PKC motif mutations are required to confirmed whether this motif is critical for this activity (Mubin et al., 2010). The conserved arginine residue present in the PKC motif of TbCSV V2 is required for RNA silencing suppression activity (Li et al., 2021a). The V2 of BCTV (genus *Curtovirus*), is functionally homologous to the V2 of begomoviruses despite

sharing low sequence similarity. Luna et al. (2020) found that mutation of threonine present in the BCTV V2 PKC motif to alanine disrupted ability of V2 to act as a silencing suppressor. In addition, the presence of the threonine was necessary to induce systemic infection and the mutant protein displayed an altered subcellular localisation (Luna et al., 2020).

Another conserved begomoviral V2 motif is the CxC motif which has a consensus sequence of Cys-x-Cys, where x is any aa residue. Interestingly, the CxC motif is absent in BCTV V2 (Luna et al., 2020). Mutation of the two conserved cysteine residues to serine in a begomovirus AV2 protein resulted in significantly reduced pathogenicity (Padidam et al., 1996). Mutation of these two residues to serine in both TYLCV and TYLCV-Is V2 proteins terminated the protein's ability to suppress RNA silencing but did not alter subcellular localisation in the case of mutant TYLCV V2 (Glick et al., 2008; Zrachya et al., 2007). A disruption in the interaction of V2 with tomato SGS3 was suggested as a mechanism behind this loss of silencing suppression (Glick et al., 2008; Zhao et al., 2020). Fukunaga and Doudna (2009) suggested that a reduction in competitive dsRNA binding activity of mutant V2 was responsible for the reduced silencing suppression rather than altered V2-SGS3 binding. Interestingly, this double mutation in the CxC motif did not reduce the ability of TYLCV V2 to suppress TGS (Wang et al., 2014b). Several additional residues have been shown to be essential for V2 protein function. Using substitution mutants of TbCSV V2, Li et al. (2021a) demonstrated that several conserved arginine, histidine, phenylalanine, and tyrosine residues are required for efficient silencing suppression activity.

1.3.3 Putative ORFs in the V1/V2 coding region

The C5/AC5 ORF is present in the V1/V2 coding region and is in frame with V1 but is present on the cs strand. This putative ORF has been previously mentioned in early studies of OW mono- and bipartite geminivirus species belonging to the genus *Begomovirus* including AYVV, *Indian cassava mosaic virus* (ICMV), and *Tomato leaf curl New Delhi virus* (ToLCNDV) (Hong et al., 1993; Padidam et al., 1995; Tan et al., 1995). These hypothetical C5/AC5 protein sequences were deemed small, rare, and not conserved among various geminiviruses. They consequently received little interest

in favour of the larger ORFs such as C1/AC1 and V1/AV1. A study of pepper huasteco yellow vein virus (PHYVV) by Torres-Pacheco et al. (1993) reported the presence of a C5 ORF (termed AL5). Multiple sequence alignments with other geminivirus AL5 showed some sequence conservation but no transcript or protein product was identified, and elements involved in the regulation of this ORF were not determined. In addition to AC5, two putative ORFs have been identified in the AV1/AV2 coding region of Hedyotis uncinella yellow mosaic virus (HeuYMV): AC6 and AV3 (Du et al., 2014). Their expression was not confirmed however, and their role in infectivity was not determined (Du et al., 2014).

Studies examining the role of C5/AC5 in pathogenicity and infection have shown contradictory results which may be expected since C5/AC5 is not a conserved ORF and may serve accessory, rather than key, functions. Site directed mutagenesis of an infectious clone of watermelon chlorotic stunt virus (WmCSV) to introduce three stop codons in the AC5 ORF encoding a 255 aa protein resulted in no change to disease symptoms or virus infectivity (Kheyr-Pour et al., 2000). Fontenelle et al. (2007) mutated the AC5 ORF (250 aa) of tomato chlorotic mottle virus (ToCMoV) and reported no change in infectivity or symptom development compared to the plants inoculated with wild-type virus. The authors did not detect an AC5 transcript in infected plants and concluded that AC5 was of no importance. A study of three isolates of a NW monopartite begomovirus, tomato leaf deformation virus (ToLDeV), concluded that C5 was not important for pathogenicity or infection (Melgarejo et al., 2013). C5 start codons were removed in infectious clones of each of the three ToLDeV isolates, only one of which resulted in a reduced symptom phenotype in *N. benthamiana*.

In some begomoviruses, the C5/AC5 has been shown to exert a critical role in disease development. It has been reported to localise in the nucleus, and transcripts have been detected in infected plants using RT-PCR. In addition, sequences upstream of the C5/AC5 ORF have shown promoter activity (Gong et al., 2021; Li et al., 2021b). The AC5 protein of mungbean yellow mosaic India virus (MYMIV) has been demonstrated to serve various functions related to inducing disease symptoms and hindering antiviral defence mechanisms (Li et al., 2015). An AC5 ORF deletion mutant had reduced infectivity, resulted in milder disease symptoms and a significant reduction in viral load in infected plants. Expression of AC5 using a PVX-based vector system established that the AC5 protein acts as a symptom determinant in N. benthamiana.

The N-terminal region of the AC5 protein (50 aa) was shown to function as a suppressor of PTGS, whilst the C-terminal region was found to suppress TGS (Li et al., 2015). The 114 aa C5 protein of Ageratum leaf curl Sichuan virus (ALCScV) was also determined to suppress PTGS activity, act as a pathogenicity determinant, and modulate levels of viral DNA in infected plants (Li et al., 2021b). Infection of plants with a C5 deletion mutant resulted in delayed symptom development and a reduction in symptom severity.

1.4 Replication-associated protein (Rep)

The Rep protein is the most important viral protein encoded by the ssDNA genome of plant-infecting geminiviruses and without it the virus cannot successfully infect plant host tissues (Rizvi et al., 2015; Saunders et al., 1991). Rep is responsible for recruiting and coordinating host polymerases and associated factors to drive the rolling circle amplification (RCA) of the geminivirus genome in the cell nucleus (Ruhel and Chakraborty, 2019; Hanley-Bowdoin et al., 2013). To carry out these necessary functions, Rep has been demonstrated to execute several enzymatic activities at various stages of replication. Upon binding to the viral genome dsDNA intermediate, Rep nicks the viral-sense strand at the Ori (endonuclease activity). By catalysing the hydrolysis of ATP (ATPase activity), Rep is then able to unwind the dsDNA (helicase activity). Following vs strand extension mediated by host DNA polymerases, Rep cleaves the newly synthesised vs strand and ligates the free ends (ligase activity) to generate a circular ssDNA (Argüello-Astorga et al., 1994; Choudhury et al., 2006; Clérot and Bernardi, 2006; Desbiez et al., 1995; Ilyina and Koonin, 1992; Koonin and Ilyina, 1992; Laufs et al., 1995a; Orozco et al., 2000; Pooggin, 2013). The role of Rep in begomovirus replication is discussed in more detail in Section 1.4.3. In addition to performing functions associated with RCA, ACMV Rep has also been implicated in modulation of gene expression by negatively regulating the expression of its own ORF AC1 as well as ORF AC4 (Haley et al., 1992; Hong and Stanley, 1995). The Rep protein also exerts an effect upon the host cell cycle to generate conditions which are conducive to successful infection and virus DNA replication (Fondong, 2013; Kittelmann et al., 2009). This is discussed in more detail in Section 1.4.4. To successfully infect a cell, begomoviruses must overcome host silencing defence

pathways (Hanley-Bowdoin et al., 2013). The ability of Rep to negatively affect these pathways is described in Section 1.4.4.

1.4.1 Domain organisation and functional motifs

Important residues implicated in the enzymatic activity of geminivirus Rep have been mapped to two distinct protein domains containing functional motifs conserved amongst diverse proteins with similar functions. These two domains include the N-terminal HUH endonuclease domain and a C-terminal ATPase/superfamily 3 (SF3) helicase domain (Desbiez et al., 1995; George et al., 2014; Orozco et al., 1997; Tarasova and Khayat, 2022). A short oligomerisation domain separating the two enzymatic domains is necessary for the oligomeric assembly of Rep monomers (Choudhury et al., 2006; Orozco et al., 1997). A fourth domain, the carboxyl-terminal (C-terminal) domain, has been identified in TYLCSV Rep but its precise function was not determined (Clérot and Bernardi, 2006). The C-terminal domain is discussed further in Chapter 3, Section 3.4.2.

The N-terminal domain is critical for the initiation of RCA to replicate the viral genome (Hanley-Bowdoin et al., 2013) and is responsible for the DNA binding, cleavage and ligation activities of the protein (Fondong, 2013). Orozco et al. (1997) showed that deletion of the first 29 residues of the N-terminal domain renders the TGMV Rep protein unable to perform DNA binding/cleavage. The authors were able to map three conserved motifs and assign predicted functions to each. The three motifs include: motif I with a consensus sequence of FLTY; motif II with a cs of HUHUUU, where U is a bulky hydrophobic aa residue; and motif III with a cs of YXXKD/E/N, where X is any aa residue. Motif I is critical to begomovirus replication as it has been shown to be associated with the DNA binding and DNA cleavage activity of Rep (Orozco and Hanley-Bowdoin, 1998). Motif II containing two conserved histidine (H) residues was proposed to function as a metal ion coordination site (Ilyina and Koonin, 1992; Orozco and Hanley-Bowdoin, 1998). A study conducted by Laufs et al. (1995b) revealed that the DNA cleavage activity of TYLCV Rep is decidedly dependent upon the presence of the metal ions Mg^{2+}, Mn^{2+}, or Ba^{2+} in solution. TGMV Rep mutants containing a disrupted motif II resulted in failure of DNA replication activity (Orozco and Hanley-Bowdoin, 1998). Motif III contains a conserved tyrosine residue which is catalytic in

nature. The tyrosine residue cleaves the viral DNA via a nucleophilic attack of the DNA phosphodiester bond at the Ori located on the stem-loop nonanucleotide structure. Rep cleaves between the 7th and 8th nucleotide (nt) of the conserved 5'-TAATATTAC-3' sequence on the vs strand (Laufs et al., 1995c). Upon cleavage, the hydroxyl group generated on the tyrosine residue forms a covalent bond with the 5' phosphoryl group of the nicked viral DNA and subsequent replication of the DNA strand occurs from the liberated 3' end (Laufs et al., 1995b,1995c; Nash et al., 2011). All three motifs are located on the functional portions of the N-terminus which are responsible for DNA binding, cleavage and ligation (Orozco and Hanley-Bowdoin, 1998).

Within the DNA binding and cleavage portion of the N-terminal domain between motifs I and II, two α-helices are predicted to occur (Orozco et al., 1997; Orozco and Hanley-Bowdoin, 1998). The amino acid (aa) sequences of these two α-helices (12 and 10 residues in TGMV Rep respectively) contain charged residues which are conserved amongst geminivirus Rep proteins (Orozco and Hanley-Bowdoin, 1998). Site-directed mutagenesis of the α-helices in TGMV Rep resulted in loss of DNA replication ability by mutant Rep. This indicated that the conserved α-helix residues are essential for the DNA binding/cleavage functionality of the Rep protein (Orozco and Hanley-Bowdoin, 1998). Recently, a fourth motif conserved across geminiviruses, termed the geminivirus replication sequence (GRS) was identified by Nash et al. (2011). The GRS motif is positioned between motifs II and III in the DNA binding/cleavage portion of the N-terminus. The ability of mutant TGMV Rep containing a modified GRS motif (alanine substitution) to bind ds/ssDNA was similar to that of the wild type Rep. DNA cleavage and replication activity was lost however, and viral genomes encoding mutant GRS motif failed to induce a TGMV-induced disease phenotype in inoculated *N. benthamiana* (Nash et al., 2011). The GRS motif is thus critical to viral genome replication.

The ability of Rep to form oligomeric structures is linked to the predicted oligomerisation domain (Clérot and Bernardi, 2006). The DNA binding activity of Rep is dependent upon the formation of Rep oligomers from translated monomers, whilst oligomerisation is not necessary for the DNA cleavage and ligation activity of Rep (Orozco et al., 1997, 1998). It has been suggested that despite being necessary for DNA binding, the oligomerisation domain itself does not make physical contact with the DNA molecule (Orozco et al., 1998). In TGMV, the oligomerisation domain of Rep

has been mapped to between residues 120 and 181, although residues flanking this region enhance the stability of Rep oligomers and promote multimerisation (Orozco et al., 2000). Two α-helices are predicted to occur between TGMV Rep oligomerisation domain residues 131 and 152, an area where there exists an overlap between the DNA binding region of Rep and the oligomerisation domain. It is likely that these helices contribute to both DNA binding and oligomerisation activities (Orozco et al., 1997). TGMV Rep deletion-mutants lacking these two α-helices showed greatly reduced oligomerisation activity and mutations introduced into the oligomerisation domain negatively affected viral DNA replication (Orozco et al., 2000). Helicase activity is reliant upon oligomerisation in the case of TYLCSV Rep (Clérot and Bernardi, 2006). The reported oligomerisation for geminivirus Rep proteins differs. TGMV Rep protein has been reported to form octamers (Orozco et al., 1997). The truncated Rep of TYLCSV was able to form oligomers varying from a hexamer to a dodecamer(double-hexamer) or larger. For helicase activity and subsequent bi-directional replication to occur, it has been predicted that Rep must be present as a dodecamer or larger (Clérot and Bernardi, 2006). Successful helicase activity of MYMIV Rep is obtained upon the assembly of Rep into a large 24-mer complex, and mutations introduced into the oligomerisation domain of MYMIV Rep resulted in disruption of helicase activity (Choudhury et al., 2006; Singh et al., 2007). The Rep protein of wheat dwarf virus (WDV) forms octamers and the assembly of monomers to form oligomers in solution is dependent upon pH. A solution pH ≤7.0 favours octamer formation whilst a pH ≥7.4 favours Rep monomers. The formation of a viral DNA-protein complex is favoured at solution pH values of 7.0-7.4. It has been suggested that WDV Rep binding to DNA occurs as monomers with subsequent Rep-Rep interaction (once DNA-bound) resulting in an oligomer (Missich et al., 2000). As mentioned previously, REn is known to bind to, and form complexes with Rep to facilitate viral genome replication (Settlage et al., 1996). The location of REn binding has been mapped to an area (aa 101-181) which encompasses the oligomerisation domain of Rep in TGMV, although the area of protein contact may be smaller. It has also been revealed that Rep oligomerisation is a likely prerequisite necessary for REn binding to occur (Settlage et al., 2001).

The ATPase/helicase domain present on the C-terminus of begomovirus Rep contains residues which are responsible for the hydrolysis of adenosine triphosphate (ATP) and guanosine triphosphate (GTP) as well as helicase activity (Clérot and Bernardi, 2006;

Desbiez et al., 1995; George et al., 2014; Rizvi et al., 2015; Vadivukarasi et al., 2007). Rep unwinds dsDNA in the 3' to 5' direction by utilising the energy released by ATP hydrolysis (Clérot and Bernardi, 2006; George et al., 2014; Hickman and Dyda, 2005). Helicases of this family are known to contain four conserved functional motifs (Hickman and Dyda, 2005; Koonin, 1993; Vadivukarasi et al., 2007). The four motifs include: Walker A (WA, P-loop) with a consensus sequence GXXXXGK(T/S), where X is any aa residue; Walker B (WB) with a consensus sequence of XXXXDD; motif B' with a cs of $KX_{3-4}GX_{7-8}K$ and C motif containing a conserved asparagine.

The WA motif of the Rep helicase region is responsible for the ATP binding, whilst the WB motif is responsible for ATP hydrolysis (Choudhury et al., 2006; George et al., 2014). Mutation of the WA and WB motifs results in loss of ATPase and helicase activity (Clérot and Bernardi, 2006; Desbiez et al. 1995; Orozco et al., 1997). The B' motif, which is located between WB and C motif, has been associated with a variety of functions in the helicase activity of SF3 Rep proteins. Residues located in this motif have been shown to bind DNA, be involved in ATP interaction, coupling of ATPase activity with DNA binding, as well as driving conformational shifts during helicase activity (George et al., 2014; Yoon-Robarts et al., 2004). The C motif is known to interact with the ATP terminal phosphate and recognise the associated unwinding of the DNA structure (Guo and Huang, 2010). In addition to the four motifs, a β-hairpin loop structure is predicted to exist adjacent to the B' motif, upstream from the C motif. It has been suggested that this loop, which contains a conserved lysine, is positioned in the oligomeric helicase structure to bind and interact with DNA during unwinding (George et al., 2014; Shen et al., 2005). It is interesting to note that an arginine finger residue, which is known to occur downstream of the C motif in other SF3 helicases and perform ATPase and conformational shift functions, is absent in begomovirus Rep proteins (Clérot and Bernardi, 2006; George et al., 2014).

1.4.2 Structure

The secondary structure (ss) of the TGMV Rep protein N-terminal domain has been predicted to contain two α-helices between residue positions 25-52. Two α-helices are also predicted to occur between TGMV Rep oligomerisation domain residues 131 and 152, an area where there exists an overlap between the DNA binding region of Rep

and the OD. (Orozco et al., 1997). The ss of the ATPase domain of TYLCSV was predicted by Clérot and Bernardi (2006) to contain four α-helices and a β-sheet with three strands. The β-sheet contains the WA and WB motifs as well as the C motif. The ATPase domain of the simian virus 40 (SV40) large tumour antigen (Tag), an SF3 helicase related to TYLCV Rep, is known to contain eight α-helices and a five-stranded β-sheet, indicating that TYLCV Rep has reduced structural elements in comparison (Clérot and Bernardi 2006). Analysis of Rep ATPase domain of tomato leaf curl Gujarat virus (ToLCGuV) showed the approximate ss composition of this domain to be: 37% α-helix, 13% β-sheet, and 50% loops (random coils) (George et al., 2014).

Although the molecular structure for some related SF3 viral helicases such as human papillomavirus (HPV) type 18 helicase (Abbate et al., 2004), adeno-associated virus 2 (AAV2) helicase (James et al., 2003), bovine papillomavirus (BPV) E1 helicase (Enemark and Joshua-Tor, 2006), and the SV40 Tag helicase domains (Gai et al., 2004) have been solved using X-ray crystallography; this is not the case for geminiviruses for which a solved crystal structure is yet to be published. The molecular structure of TYLCSV Rep N-terminal domain (aa 4-121) was solved using heteronuclear multidimensional nuclear magnetic resonance (NMR) spectroscopy (Campos-Olivas et al., 2002). The ss and tertiary structure of the solved structure is shown in Fig. 1.2.

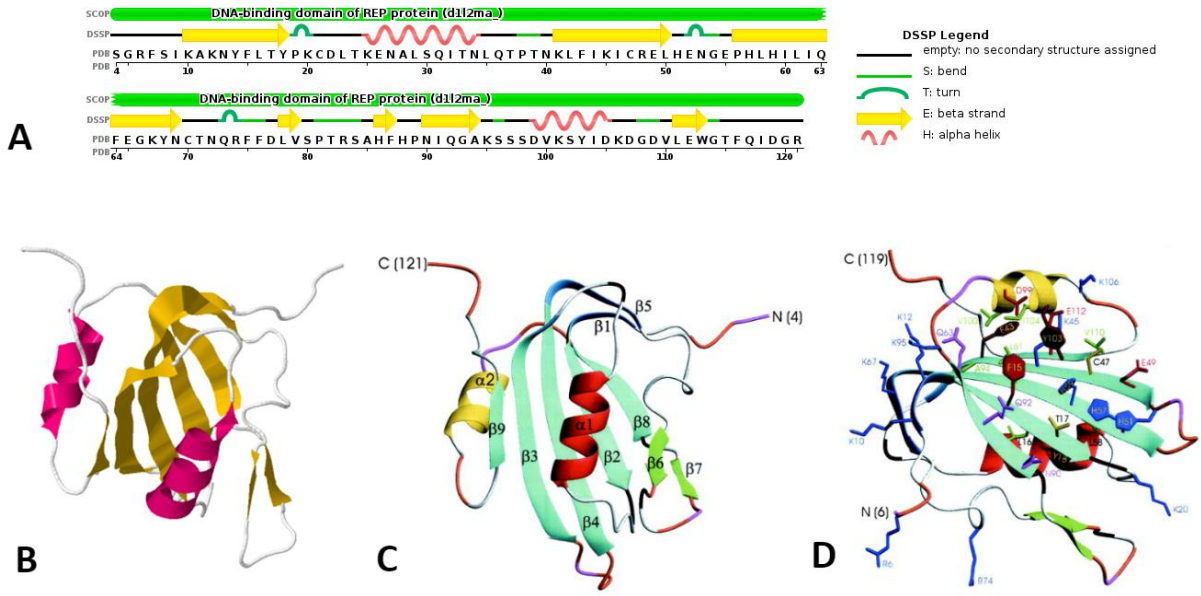


Fig. 1.2 TYLCSV Rep N-terminal domain (PDB ID 1L2M) **A**, secondary structure; **B**, three-dimensional ribbon diagram of regularised average tertiary structure showing α-helices (magenta) and β-sheets (yellow); **C**, α-helices, β-sheets, N and C termini labelled; **D**, selected aa side chains belonging to

conserved catalytic domain motifs or occupying space equivalent to the ss/dsDNA binding residues of related protein structures (from Campos-Olivas et al., 2002).

The TYLCSV Rep N-terminal domain structure contains two α-helices, nine β-strands, and several loops. A major antiparallel β-sheet contains five β-strands (β2-4, β8, β9), with the minor β-sheet containing two strands (β1 and β5). The minor β-sheet strands extend from β2 and β4. A β-hairpin loop containing two strands (β6 and β7) is present downstream of the β5 strand of the minor sheet. One face of the major sheet is largely exposed, with only the shorter α-helix flanked by partially disordered loops being present on this face. Motif I was located on β2, motif II on β4 (adjacent to β2), and motif III containing the catalytic tyrosine residue was situated on the shorter α-helix. The tyrosine residue was directed towards the exposed face of the β-sheet near motifs I and II. This close arrangement of the three motifs on the exposed face of the major β-sheet suggests that this could be the catalytic area for the activity of the N-terminal domain of Rep (Campos-Olivas et al., 2002). Recently, the crystal structure of WDV Rep N-terminal domain was solved and displays a similar structural arrangement to TYLCSV Rep (Everett et al., 2019).

In silico homology modelling was used by George et al. (2014) to predict how the monomeric structures of ToLCGuV Rep helicase domain assemble to form the quaternary hexameric helicase structure (Fig. 1.3).

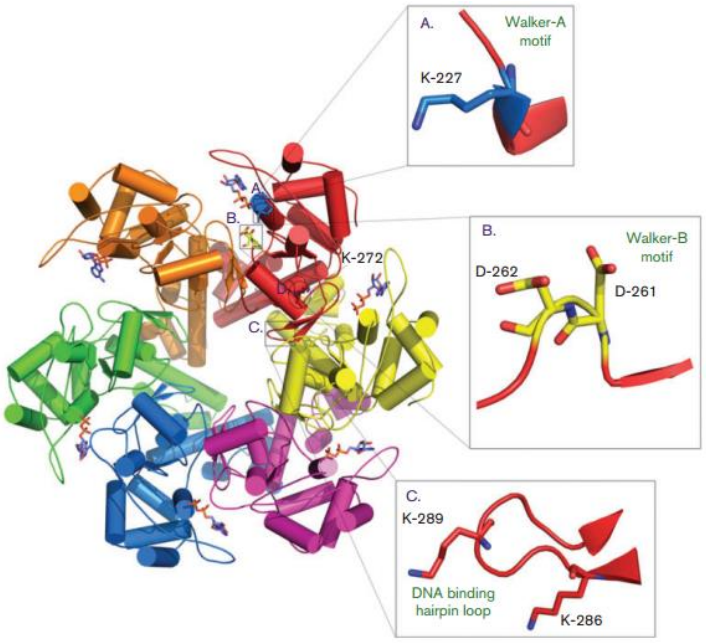


Fig. 1.3 Three-dimensional homology model of ToLCGuV Rep ATPase/helicase domain hexamer modelled using the solved crystal structure of BPV E1 helicase domain (PDB ID: 2GXA) as the template

(Enemark and Joshua-Tor, 2006; George et al., 2014). Hexamer shown as a cartoon with each monomer subunit depicted using different colours, Important Walker A motif, Walker B motif, and DNA binding hairpin loop residues shown as sticks with magnified views of the motifs depicted in boxes A-C. Adenosine diphosphate (ADP) shown as sticks coupled to each monomer (from George et al., 2014).

This was achieved by using the solved crystal structure of the ATPase domain of the BPV E1 protein (Enemark and Joshua-Tor, 2006). Despite the low aa sequence identity (16%) between E1 and ToLCGuV Rep, the authors justified the use of E1 as the template due to favourable biochemical data obtained for ToLCGuV Rep which compares appropriately given the predicted functions of the WA/WB motifs and the DNA binding loop present adjacent to motif B'. The positioning of the DNA binding hairpin loop in the predicted hexamer is such that it projects into the central channel which places it in a favourable conformation for DNA binding as predicted in other SF3 helicases (George et al., 2014). Homology modelling of ToLCNDV Rep protein was done and the interactions of several important B' motif and β-hairpin loop residues with ssDNA and ADP were investigated using molecular dynamics simulations (Ruhel et al., 2021).

1.4.3 Role in geminivirus replication

Geminiviruses replicate their DNA in host cells using RCA and since neither the Rep nor any other viral protein has polymerase activity, the virus must recruit host polymerases and associated replication factors to perform this function. A simplified model of geminivirus replication is shown in Fig. 1.4. As mentioned previously, the various conserved domains of Rep have critical functions associated with DNA replication (initiation, elongation and termination) and as a result, this protein is predicted to have an important function in assembling the viral replisome. The replisome is a closely associated complex of biomolecules and contains the host factors and viral proteins required for DNA replication (Hanley-Bowdoin et al., 2013). Once viral ssDNA enters the host cell nucleus, a complementary strand is synthesised using a host DNA polymerase resulting in a dsDNA intermediate (Gutierrez, 2000; Hanley-Bowdoin et al., 2013; Pooggin, 2013). Host RNA polymerase II transcribes viral mRNA molecules utilising the bidirectional transcription method to obtain transcripts for proteins with ORFs on the vs and cs strands (Hanley-Bowdoin et al., 2013; Pooggin, 2013).

To initiate RCA, Rep must generate a nick in the vs strand of the dsDNA intermediate (Koonin and Ilyina, 1992). Rep binding occurs at Rep-binding sites (iteron sequences) present adjacent to the Rep TATA box promoter. The stem-loop structure of begomoviruses contains the conserved 5'-TAATATTAC-3' sequence on the vs strand and Rep cleaves between the 7th and 8th nt of this sequence. (Hanley-Bowdoin et al., 2000; Hidayat et al., 2008; Rizvi et al., 2015).

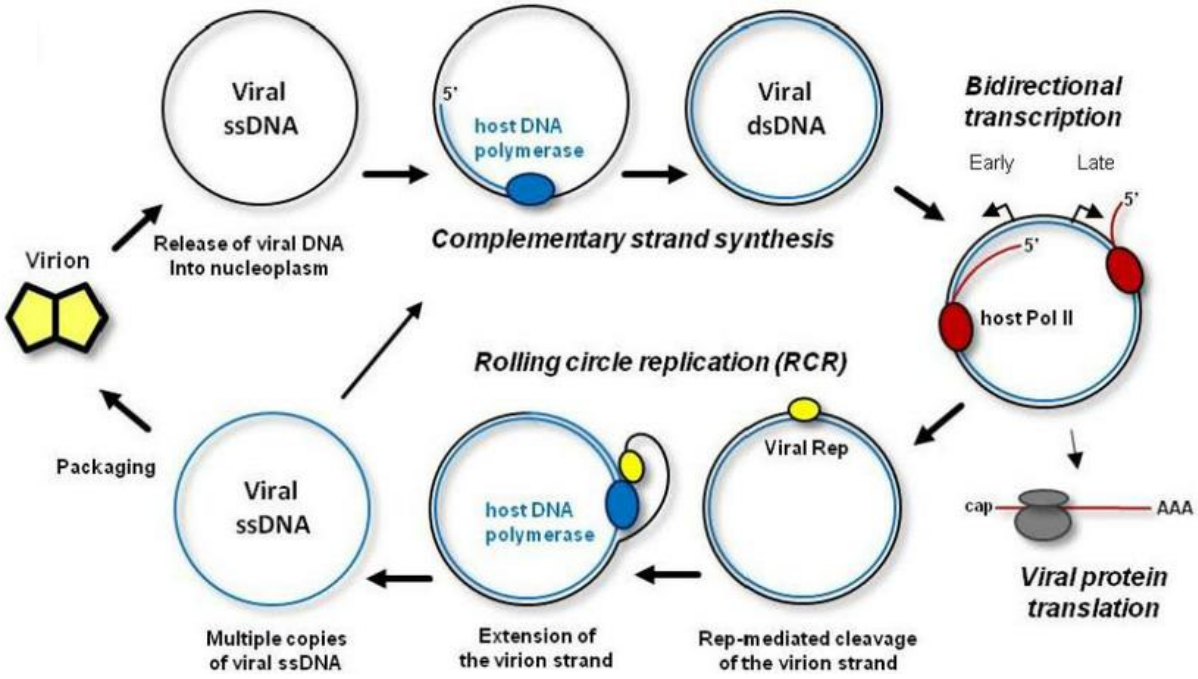


Fig. 1.4 Simplified model of the replication cycle of geminiviruses. Upon liberation from the capsid, the ssDNA entering the nucleus, a complementary strand is synthesized using host DNA polymerase to generate a dsDNA intermediate. Bidirectional transcription of viral mRNA using host RNA polymerase II results in the production of mRNA transcripts which are translated by the host creating viral protein products. Rep binds to the virion-sense strand of the dsDNA intermediate, initiating strand cleavage. Rep then facilitates rolling circle amplification (RCA) by extending the virion strand with the use of host DNA polymerases. Rep cleaves the generated virion-sense strand and ligates it to generate a circular ssDNA molecule which may undergo encapsidation or complementary strand synthesis (from Pooggin, 2013).

Proteins that perform helicase and replication-associated activities are known to bind to cruciform DNA structures which form in areas of dsDNA containing inverted repeats. The inverted repeat sequences present on one strand bind in a complementary fashion to each other forming the stem, and the nt present in the gap between the inverted repeats form the loop structure at the terminus of the stem. In dsDNA, the both strands form the stem-loop structure which results in the cross-like (cruciform)

shape (Brázda et al., 2011). Upon binding to the iteron sequence on the vs strand, WDV Rep is predicted to oligomerise (Gutierrez et al., 2004; Missich et al., 2000). In MYMIV, Rep binds in a sequence-specific manner to iteron sequences on either side of the Ori (Singh et al., 2007, 2008). Once bound to the iterons, Rep then denatures the DNA and allows for the extrusion of the cruciform structure on the vs strand. The stem-loop structure is formed and Rep is able to subsequently cleave Ori. Once Ori is cleaved, Rep binds covalently with the 5' phosphoryl group of the viral DNA and the viral replisome begins replication from the liberated 3' end (Laufs et al., 1995b, 1995c; Nash et al., 2011; Singh et al., 2007, 2008). A predicted model of the replication initiation activity of Rep is shown in Fig. 1.5.

The 3' end of the nicked DNA acts as the primer for the generation of the new vs strand which displaces the parent strand from the complementary sense strand as host polymerases generate the new strand. Once the new vs strand is created, the Ori is regenerated and Rep can continue synthesizing new strands whereas the nick in the displaced strand is ligated by Rep and the circular ssDNA may then be used as a template for complementary strand synthesis, undergo encapsidation by associating with the CP or interact with NSP/MP for subsequent viral DNA movement to occur (Gutierrez, 2000; Hanley-Bowdoin et al., 2013).

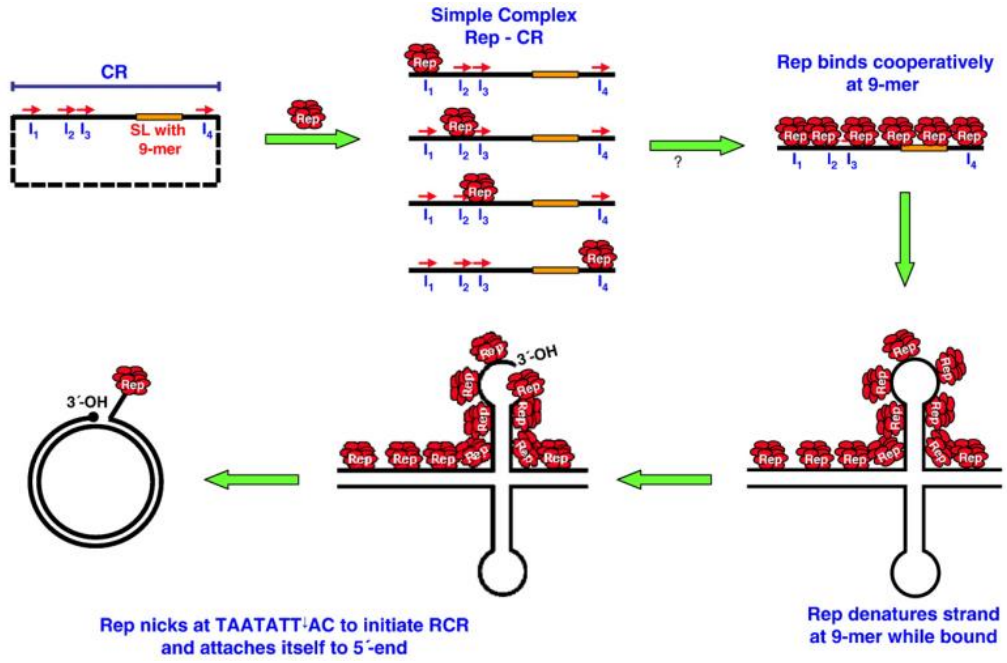


Fig. 1.5 Predicted model of the predicted replication initiation activity of MYMIV Rep. Upon binding to the iteron (I) sequences around the stem-loop (SL) structure containing the nonanucleotide (9-mer) sequence present in the common region (CR) of the viral DNA, Rep induces the formation of the

cruciform DNA structure. Rep induces a nick at the origin of replication and bonds to the 5' end, allowing replication to commence from the 3' end of the cleaved strand (from Singh et al., 2008).

1.4.4 Role in host cell cycle regulation and defence pathway silencing

As mentioned previously, viral genome replication relies on the presence of the viral replisome at the replication fork (Clérot and Bernardi, 2006). It is highly probable that the geminivirus REn protein is involved in the replisome as it has been shown to enhance viral DNA replication by oligomerising and interacting with Rep (Fondong, 2013; Pasumarthy et al., 2010; Settlage et al., 1996). It has been shown that TYLCSV REn, in addition to interacting with Rep, is able to undergo interaction with plant homologues of the retinoblastoma protein (pRb) as well as PCNA (Castillo et al., 2003; Settlage et al., 2001). The host protein PCNA is a critical component of the plant cell replication machinery which begomoviruses use to replicate in the host. PCNA is known to function in eukaryote DNA chromosome replication whilst pRb is a critical cell cycle regulator (Castillo et al., 2003; Kong et al., 2000). Rizvi et al. (2015) hypothesized that in addition to the viral Rep and REn proteins of the replisome, host proteins forming part of the replication machinery include PCNA, replication factor C protein (RFCp), DNA polymerase, replication protein A32 (RPA32) and minichromosome maintenance 2 (MCM2).

Geminiviruses are known to infect mature cells which are not undergoing an active cell cycle and thus do not contain replication machinery or active DNA polymerases (Hanley-Bowdoin et al., 2013). To overcome this and allow for the replication machinery to assemble in the infected cell, Rep binds to pRb. When pRb is bound to the host E2F transcription factor it results in the repression of genes which encode the proteins associated with replication machinery. When Rep in association with REn binds pRb, it inactivates the repression of replication factors such as PCNA and this allows the cell to enter S-phase which is when the cell is in a state of active replication (Argüello-Astorga et al., 2004; Castillo et al., 2003; Egelkrout et al., 2001; Hanley-Bowdoin et al., 2013; Kong et al., 2000).

When viral DNA is converted to its dsDNA replicative form, the DNA molecules associate with host histone proteins to form a minichromosome which contains up to 12 nucleosomes. Rep is known to bind with host histone 3 protein and MCM2. It has been hypothesized that this interaction is critical for the dynamic interactions and processes occurring at the minichromosome (Hanley-Bowdoin et al., 2013). The Rep protein of chilli leaf curl virus (ChiLCV) has been found to interact with plant homologues of enzymes present in the monoubiquitination machinery to promote and enhance viral gene transcription activity in the minichromosome (Kushwaha et al., 2017).

Rep also interacts with the large subunit of the RFCp complex which is essential for the functioning of PCNA since RFCp acts as a clamp loader that allows PCNA to associate with DNA (Castillo et al., 2003; Luque et al., 2002). The Rep protein of MYMIV has been shown to interact with the middle subunit of the host replication factor A and this interaction was found to result in improved function of the Rep ATPase domain (Singh et al., 2007) as well as providing stability to newly formed ssDNA during replication (Singh et al., 2008). Whilst Kaliappan et al. (2012) claimed that the host DNA repair and recombination protein (RAD54) was critical to the rolling circle replication of MYMIV, Richter et al. (2015) disputed their findings and concluded that RAD54 was not essential for the replication of two geminiviruses via complementary strand, rolling circle or recombination-dependent replication.

When expressed in yeast and insect cells, the oligomerisation domain of geminivirus Rep binds and interacts with two host proteins: geminivirus rep-interacting kinase (GRIK) and geminivirus rep-interacting motor protein (GRIMP). The interaction is thought to enhance the conditions present in the cell to allow for viral replication to occur. Rep was also found to interact with host histone H3 with the possible outcome of direct improvement of viral replication activity (Kong and Hanley-Bowdoin, 2002). Another interaction which is hypothesized to improve the cellular environment for viral replication is that of Rep with another host factor termed the small ubiquitin-related modifier (SUMO)-conjugating enzyme (SCE1) (Kushwaha et al., 2017). The improvement arises when the interaction between Rep and SCE1 adjusts the post-translational modification (SUMOylation) of several host targets (Kushwaha et al., 2017; Sánchez-Durán et al., 2011). The binding sites for pRb, PCNA and SUMO are located in the vicinity of the junction between the N-terminal domain and the

oligomerisation domain (Fondong, 2013). It is evident that geminivirus Rep interacts with a variety of host factors to not only promote active replication, but to also facilitate a suitable cell environment for viral DNA replication to occur.

The bipartite geminivirus TrAP and AC4 proteins as well as the monopartite geminivirus V2, TrAP and C4 proteins have been implicated in the suppression of RNA silencing pathways in the plant host (Bar-Ziv et al., 2015; Castillo-Gonzáles et al., 2015; Chellappan et al., 2005; Glick et al., 2008; Hanley-Bowdoin et al., 2013; Liu et al., 2017; Melgarejo et al., 2013; Peretz et al., 2011; Trinks et al., 2005). It was reported that geminivirus Rep acts as a suppressor of TGS in *Arabidopsis thaliana* by interfering with DNA (cytosine) methylation (Rodríguez-Negrete et al., 2013). The authors showed that Rep can repress two host factors necessary for DNA methylation to occur namely chromomethylase 3 (CMT3) and MET1. The repression of MET1 by Rep is significantly improved by the presence of geminivirus C4 protein, whilst the repression of CMT3 is not dependent on the presence of the C4 protein (Rodríguez-Negrete et al., 2013).

1.5 Rationale for study

Tomato is an important crop to both the commercial and subsistence farming industries in Southern Africa. Recombination events occurring within the dynamic populations of begomoviruses in mixed infection drives their evolution. This combined with the increased prevalence of the polyphagous whitefly vector means that the potential for begomovirus diseases of crop plants to worsen in the future is high. Surveys of ToCSV in Southern Africa indicate that this virus is undergoing evolution and therefore additional research will prove helpful to mitigate the impacts of this pathogen (Esterhuizen, 2012; Sande et al., 2021). The recent discovery of highly divergent geminiviruses in wild plant and weed hosts in SA is particularly concerning because it suggests that significant genetic diversity exists in the field. Such diversity, although fascinating from the geminivirologist's perspective, could prove disastrous for the developing African agricultural industry. Serious crop losses could occur with the emergence of a novel begomovirus in local crop cultivars.

Concurrent with the growing number of recognised geminivirus species, information regarding the role of several important geminivirus proteins in viral infection has

increased significantly. Specific sequence differences between geminiviruses in the V2, IR and other genomic regions contribute to their highly diverse pathologies. The existence of newly recognised coding regions and promoter elements highlight that much remains to be discovered regarding the pathways involved in disease development. Geminivirus Rep is one of the most important viral proteins which modifies the host cell environment and manipulates the host cell machinery to allow for viral replication to occur. It is therefore imperative to better understand the structure and functionality of Rep.

1.5.1 Experimental novelty

A previous study of the genetic diversity of ToCSV in SA resulted in the identification of two novel variant groups (Esterhuizen, 2012). Although these two variant groups were shown to induce disease in tomato at significantly different levels of severity, the influence of specific sequence differences on pathogenicity was not investigated in detail. Infectivity trials of ToCSV in *N. benthamiana* have not been conducted previously and the symptom phenotype induced by representative isolates from these two groups has not been reported prior to the current study. This host was selected because it is used extensively to characterise plant viruses and investigate their pathogenicity in vivo. Its increased susceptibility is believed to be as a result of the presence of a dysfunctional RDR gene in the wild type genome which contains in-frame stop codons in the 5' gene region (Goodin et al., 2008; Yang et al., 2004).

Studies of putative ORFs in ToCSV and related African begomovirus species are lacking and the present study serves as a ground-breaking effort to improve the awareness of rare and unusual putative ORFs in the DNA-A of monopartite and bipartite begomoviruses. With these knowledge gaps in mind, the recent publication of several important findings related to novel sequence features of geminiviruses (Cantú-Iris et al., 2019; Gong et al., 2021) have served as a motivating factor to investigate the differences between ToCSV variant sequences and related African begomovirus isolates. This was done with the hope of better understanding these important pathogens. Except for the solved structures of the N-terminal endonuclease domain of geminivirus Rep (Campos-Olivas et al., 2002; Everett et al., 2019), there are no other published structures of geminivirus Rep proteins. The tertiary structure of

ToCSV Rep has not been previously investigated and no reports of recombinant expression of ToCSV Rep ATPase/helicase domain in *Escherichia coli* are available. Whilst other studies investigating the role of specific sequence motif residues on Rep function have been published (George et al., 2014; Ruhel et al., 2021) solving the structure of this important domain remains a challenge and efforts to solve the structure will add greatly to the knowledgebase of SF3 helicases.

1.5.2 Study aims

The primary aims of this study are as follows:

1. Examine the coding capacity of ToCSV and related African begomoviruses using in silico methods.
2. Record the disease pathologies of two ToCSV variant isolate infectious clones in *N. benthamiana* and investigate the role of specific sequences in observed differences in infectivity and/or symptom phenotype.
3. Analyse the structure and function of ToCSV Rep using in silico and lab-based methods.

1.5.3 Specific study objectives

- Identify ORFs encoding proteins ≥40 aa predicted to occur in the DNA-A sequences of two ToCSV variant isolates (V30 and V22) and related African begomoviruses using ORF finder.
- Conduct canonical and putative virus protein multiple aa sequence alignments to identify regions of conservation and investigate sequence diversity using phylogenetic analysis.
- Agroinoculate *N. benthamiana* separately with two ToCSV variant infectious clones, V30 IR swap and V30/V22 V2 ORF swap mutant chimeras, and a V30 C6 deletion mutant.
- Conduct disease symptom scoring using a scoring system developed in this study and measure viral titre using qPCR at specific timepoints.
- Probe for RNA transcripts derived from selected regions of the V30 and V22 viral genomes present in infected *N. benthamiana* using RT-PCR

- Determine the aa sequence positions of endonuclease and ATPase/helicase domain functional motifs for the Rep proteins of ToCSV and related African begomoviruses.
- Use in silico homology modelling to investigate the predicted structure of the endonuclease and ATPase/helicase domains of ToCSV Rep and compare with solved structures of related virus Rep proteins.
- Examine the function of the Rep C-terminal domain in ToCSV infection by generating C1 deletion mutants using site-directed mutagenesis and conducting infectivity trials in *N. benthamiana* with symptom and viral load assessment.
- Recombinantly express partial ToCSV Rep (C-terminal region) using three expression vectors in three *E. coli* expression strains.
- Purify recombinantly expressed protein using liquid chromatography and conduct tag cleavage trials.

1.5.4 Brief outline of the thesis

Using agroinfectious clones of two ToCSV variant isolates generated previously (Esterhuizen 2012), the investigation of their coding capacity in silico and their pathology in *N. benthamiana* with additional virus mutant chimeras is presented in Chapter 2. A study in silico of ToCSV Rep, in vivo assessment of C1 deletion mutant infectivity in *N. benthamiana* and recombinant expression of the C-terminal domains of ToCSV Rep is described in Chapter 3. The concluding remarks for this study are presented in Chapter 4.

Chapter 2: Sequence differences present in the region spanning the 3' IR and the 5' V2 ORF are responsible for altered symptom phenotypes in infected *Nicotiana benthamiana*

2.1 Introduction

Tomato (*Solanum lycopersicum* L.) is a versatile agricultural food crop belonging to the dicotyledonous plant family Solanaceae and is grown exclusively for its fruit which is considered a culinary vegetable (Naika et al., 2005). It ranks second in economic importance as a solanaceous crop after potato (Simonne et al., 2011). The nutritious tomato fruit is rich in minerals, vitamins, and carotenoids (Raiola et al., 2014). In SA, the commercial industry is composed of nearly 700 producers and as of 2011, supplied 95 % of the total annual fruit output and employed over 22,000 people (DAFF, 2019). As tomato is largely harvested by hand, the industry provides employment opportunities for low-skilled agricultural labourers (Heuvelink and Dorais, 2005).

Tomato is highly susceptible to infection by several viruses transmitted by the whitefly *Bemisia tabaci*. A recent survey of tomato-growing regions in SA revealed that infestations of this polyphagous insect pest are widespread and particularly severe in the Limpopo province which is a major tomato producing region (Moodley et al., 2019). TYLCV and an ever-expanding group of related whitefly-transmitted viruses having circular ssDNA genomes belonging to the genus *Begomovirus* (family *Geminiviridae*) continue to cause devastating losses to the tomato crop worldwide (Czosnek et al., 2002; Mabvakure et al., 2016; Yan et al., 2021). Both mono- and bipartite begomoviruses contain a circular DNA-A component (~2.6 kb). Bipartite begomoviruses contain a second genomic component, termed DNA-B (~2.6 kb) (Zerbini et al., 2017). ToCSV belonging to the genus *Begomovirus* is related to the TYLCV-related complex of tomato-infecting begomoviruses and was first detected in the Onderberg region of Mpumalanga Province (Pietersen et al., 2000). ToCSV can induce up to 100% yield losses in susceptible cultivars (Pietersen and Smith, 2002). ToCSV induces similar symptoms to TYLCV in tomato, including yellowing of the young upper leaves, curling and cupping of leaflet margins, severe stunting and dwarfing of overall plant growth as well as decreased fruit set (Pietersen et al., 2000, 2008). A study of several tomato cultivars which are known to show resistance to

TYLCV were found to also show resistance to ToCSV but none of the cultivars were immune to ToCSV infection (Pietersen and Smith, 2002).

The genome of ToCSV contains six recognised ORFs found in other monopartite OW begomoviruses (Fiallo-Olivé et al., 2021; Pietersen et al., 2000; Zerbini et al., 2017). Two are located on the vs strand: V1 encoding the CP and V2 encoding the pre-coat protein (V2). Four are located on the cs strand: C1 encoding Rep, C2 encoding the transcriptional activator (TrAP), C3 encoding REn, and C4 encoding the C4 protein (Zerbini et al., 2017). The IR present between the C1 and V2 ORFs contains a stem-loop element containing a conserved nonanucleotide sequence (5'-TAATATTAC-3') at the Ori (Fiallo-Olivé et al., 2021; Piedra-Aguilera et al., 2019). The IR contains several promoters and *cis*-acting elements on either side of the stem-loop that are involved in regulating the expression of vs and cs ORFs (Ashraf et al., 2014; Borah et al., 2016; Kumar and Shivaprasad, 2020). The IR also contains repeat sequence (iteron) motifs present in the 5' IR situated in the vicinity of the TATA box core promoter sequence which act as binding sites for Rep which binds viral DNA to initiate cleavage at the Ori of the vs strand (Argüello-Astorga et al., 1994; Fontes et al., 1994).

The 3' end of the IR terminates at the start codon for the V2/AV2 ORF which is present in OW begomoviruses only (Fiallo-Olivé et al., 2021). The V2/AV2 protein is known to play a key role in virus pathogenicity and exhibits a multifunctional nature that contributes to virus-host interactions. Early studies of monopartite geminiviruses provided evidence that an intact V2 protein is required for successful systemic infection, implicating it in facilitating movement within the host (Padidam et al., 1996, Rigden et al., 1993; Stanley et al., 1992; Wartig et al. 1997). PVX-vector based expression of several geminivirus V2/AV2 ORFs in *N. benthamiana* has demonstrated that not only is V2/AV2 able to induce severe virus-like symptoms such as leaf curling and chlorosis but has also been implicated to function as an avirulence factor that promotes the development of a HR (Luna et al., 2020; Mubin et al., 2010; Saeed et al., 2018; Sharma and Ikegami, 2010). Several geminivirus V2/AV2 proteins have been demonstrated to function as pathogenicity determinants necessary for inducing severe disease (Amin et al., 2011; Chowda-Reddy et al., 2008; Luna et al., 2020; Mubin et al., 2010; Roshan et al., 2018; Sharma and Ikegami, 2010; Sharma et al., 2010; Zhang et al., 2012). In addition, V2/AV2 proteins have been reported to act as potent suppressors of TGS and PTGS by several modes of action. These include

competitive binding with important host proteins involved in antiviral defence, dsRNA binding to prevent siRNA biogenesis, as well as siRNA sequestration activity (Glick et al., 2008; Fukunaga and Doudna, 2009; Mubin et al., 2019; Wang et al., 2014b, 2018, 2019, 2020; Zhang et al., 2012).

A survey of tomato-growing regions in SA resulted in the discovery of two variant groups of ToCSV (ToCSV-I and ToCSV-II), differentiated by the presence of a recombinant fragment present in the coding region spanning part of the 3' IR and the first half of the V2 ORF in ToCSV-II isolates (Esterhuizen, 2012). In the present study the DNA-A coding capacity and in vivo infectivity of single representative isolate DNA-A genome sequences from both variant groups was investigated. Three putative cs ORFs (C5, C6, and C7) and two putative vs ORFs (V3 and V4) were identified. The C5/AC5 and C6/AC6 proteins of African begomoviruses were described in detail, conserved sequence domains were annotated and a nomenclature for these two ORFs based on their positions relative to each other and to the overlapping canonical vs ORFs was proposed. Through generation of virus mutant chimeras, we identified key sequence differences present in the IR and V2 which modulate symptom phenotype, viral titre, and disease recovery in infected *N. benthamiana*. We also investigated the role of the putative V30 C6 C-terminal region in virus infectivity and present evidence for the existence of cs RNA transcripts derived from the C6 ORF.

2.2. Methods and Materials

2.2.1 ToCSV infectious clone constructs

Agroinfectious clone constructs of the wild type severe ToCSV variant isolate ToCSV-[ZA:Mks30:08] DNA-A (OK813888) and mild ToCSV variant isolate ToCSV-[ZA:Mks22:07] DNA-A (OK813889) in pCAMBIA2300 were constructed previously (Esterhuizen, 2012). The severe and mild variant ToCSV infectious clone constructs are herein referred to as "V30" and "V22" respectively. The nt sequences of both V30 and V22 ToCSV infectious clones are included in Appendix (A) 1. Three partial IR sequence swap mutants and two full V2 ORF sequence swap mutants were synthesized. The templates used and regions swapped to generate each mutant chimera are described in Supplementary 2 (S2) Table 2.1. The V30ΔIR-s, V30ΔV2-s, and Δ22ΔV2-s mutants were synthesised by Genscript Biotech (Piscataway, USA).

The V30ΔIR-sl and V30ΔIR-sr mutants were synthesised by Inqaba Biotech (Pretoria, SA). The identities of the mutated regions of the swap mutant clones were confirmed using PCR with Sanger sequencing. Sequence chromatograms for V30 IR swap mutants are shown in A2 Fig. 1. Details of sequencing primers used are indicated in S2 Table 2.2.

2.2.2 Phylogenetic and amino acid sequence analyses

Full genome nt sequences of ToCSV-[ZA:Mks30:08] and ToCSV-[ZA:Mks22:07] were aligned with selected African tomato-infecting begomoviruses using MUSCLE alignment program with MEGA 6.06 software. A phylogenetic tree of full-length DNA-A sequences was constructed using the maximum-likelihood method (1,000 replications). A pairwise nt sequence identity matrix (MUSCLE alignment) was generated using SDT v 1.2 software. A multiple nt sequence alignment of the conserved sequence elements present in the IR spanning the stem-loop to the V2 start codon of the two ToCSV variant sequences as well as selected African begomoviruses sharing ≥78 % DNA-A sequence identity with ToCSV was constructed manually. The ORF finder server (https://www.ncbi.nlm.nih.gov/orffinder/) was used to identify ORFs predicted to encode proteins ≥40 aa in length in both wild type ToCSV variants when following standard code.

All wild type canonical and putative ToCSV protein sequences were analysed for predicted nuclear localisation signals (NLS) using cNLS Mapper (https://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi) (Kosugi et al., 2009), as well as chloroplast transit peptides (cTP) and mitochondrial transit peptides (mTP) using TargetP – 2.0 (https://services.healthtech.dtu.dk/service.php?TargetP) (Armenteros et al., 2019). Signal peptides (SP) were predicted using TargetP – 2.0 and Phobius (https://phobius.sbc.su.se/) (Käll et al., 2007). Transmembrane helices (TMH) were predicted using Phobius and TMHMM v 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) (Krogh et al., 2001). ORFs encoding V2 as well as putative V3/AV3, V4/AV4, C5/AC5, C6/AC6, and C7 proteins were identified for both ToCSV variants as well as selected African mono- and bipartite begomoviruses sharing ≥75% DNA-A sequence identity with ToCSV using BLAST analysis (Altschul et al., 1990).

Full length V2, V3/AV3, V4/AV4, C5/AC5, C6/AC6 and C7 protein sequence alignments were done using MUSCLE alignment program with MEGA 6.06 software and phylogenetic trees were constructed using the neighbour-joining method (1,000 replications). Predicted aa sequences were also aligned using PRALINE multiple sequence alignment server (https://www.ibi.vu.nl/programs/pralinewww/) (Simossis and Heringa, 2005). The ss was predicted using PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) (Buchan and Jones, 2019; Jones, 1999). Sequence logos were generated for selected protein regions using the WebLogo v 2.8.2 server (https://weblogo.berkeley.edu/logo.cgi) (Crooks et al., 2004). Sequences were manually annotated, and consensus strand and helix ss was indicated using cartoons above sequence alignments. Details of all viral isolates used in analyses are shown in S2 Table 2.3.

2.2.3 Site directed mutagenesis

ToCSV V30 was used as a template to obtain a putative V30 C6 ORF deletion mutant by replacing V30 C6 Lys109 codon with a stop codon to truncate the C-terminus by 35 aa. The deletion mutant (V30ΔC6-d) was generated via site directed mutagenesis using Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher Scientific). A partially overlapping mutagenesis primer pair was manually designed following published recommendations with amendments (Laible and Boonrod, 2009; Liu and Naismith, 2008; Zheng et al., 2004). The target mutation was included in both primers, Tm differences for the primer pair did not exceed 2 °C, and both primers had a GC content not exceeding 52 %. The OligoAnalyzer™ tool (https://www.idtdna.com/oligoanalyzer) was used to determine predicted hairpin, homo- and heterodimer ΔG values. The accepted cut-offs for predicted ΔG values were -2.0 kcal/mol for hairpins, and -9.0 kcal/mol for dimers. Primers were manually adjusted until acceptable ΔG values were obtained whilst following strict published mutagenesis primer requirements. All mutagenesis primers were synthesised, and RP-cartridge purified by Inqaba Biotech (Pretoria, SA). Mutagenesis PCR was set up as per manufacturer's instructions using 4 ng template DNA with 1 % DMSO being included to inhibit ss formation. A three-step cycling protocol was followed with an annealing temperature of 69 °C and an extension time of 520 s (45 s/kb) for 20 cycles. Upon completion, 5 µL samples were resolved on 0.7 % agarose to verify the presence of the target product at ~12 kb.

Restriction digestion using FastDigest DpnI (Thermo Fisher Scientific) was done by adding 1 µL enzyme directly to each of the remaining PCR product samples and incubating at 37 °C for 15 h followed by heat inactivation. Preparation and transformation of ultracompetent *Escherichia coli* XL10-Gold cells was done as per Green and Sambrook (2020) with the following modifications: 2 uL of DpnI-digested sample was added directly to 100 µL cells and β-mercaptoethanol (24 mM final concentration) was included to enhance transformation efficiency. Transformed cells were plated on SOB agar containing 100 µg/mL kanamycin and incubated at 37 °C for 24 h. Three to five colonies were selected and sub-cultured, plasmid DNA was extracted and quantified. All mutant clones generated in this study were confirmed using PCR (V2-specific primers) with Sanger sequencing. ORF finder was used to confirm that no inadvertent mutations had been introduced into the overlapping ORFs. Sequence chromatogram for V30ΔC6-d is shown in A2 Fig. 2. Details of mutagenesis and sequencing primers used are indicated in S2 Table 2.2.

2.2.4 Plant material and growth conditions

Wild type *N. benthamiana* was seeded in Jiffy® 7C cocopeat pellets in a controlled environment facility with a 16 h light/8 h dark photoperiod, relative humidity ~50 % and an ambient temperature of 26 ± 2 °C. Germination was induced under clingwrap and 30-day-old seedlings were acclimatised over a period of 4 to 7 days prior to inoculation at the 5 to 6-leaf stage. Plants were transferred in Jiffy® pellets to pre-wet cocopeat four days post inoculation (dpi). Supplementation with 20 mL per plant with Kelpak seaweed extract was done at 4 dpi and Seagro organic fertiliser at 8, 16, 26, and 36 dpi.

2.2.5 Inoculum preparation and plant agroinoculation

Competent *Agrobacterium tumefaciens* C58C1 cells were prepared and subsequently transformed with purified infectious clone construct plasmid DNA in accordance with Höfgen and Willmitzer (1988). Transformed cells were plated on LB agar containing 100 µg/mL rifampicin and 100 µg/mL kanamycin and incubated at 28 °C for 48 h. Positive transformants were identified and subcultured in LB broth with antibiotics until OD_{600} ~1.0 was obtained. To prepare inoculum, cultures were centrifuged and

resuspended in chilled sterile 0.1 M $MgCl_2$ containing 100 µM acetosyringone to OD_{600} 0.8. Cell suspensions were incubated in the dark at ambient temperature for up to 3 h with occasional gentle shaking. Plants were inoculated with 100 µL prepared inoculum using a 1 mL needleless syringe, with gentle infiltration of the underside of two expanded leaves below the meristem. Five plants were inoculated per treatment. Plants inoculated with *A. tumefaciens* C58C1 harbouring empty pCAMBIA2300 vector served as the mock-inoculated controls, non-inoculated controls were also included.

Preliminary infections of *N. benthamiana* with wild type ToCSV infectious clone constructs were conducted to develop a novel symptom severity scoring system. Two symptoms of interest were identified: upward leaf roll (ULR), and swollen veins (SV). For scoring of ULR symptoms, the leaf edges were demarcated into seven zones (Fig. 2.1). A visual scoring system for each symptom was developed by ranking symptom severity on a scale of zero to three, with a higher score corresponding to increased severity (Table 2.1 and Table 2.2). All inoculated plants were observed daily for symptom development. Symptom scoring was done at four-day intervals starting at 12 dpi; three expanded leaves below the meristem were scored for each plant. Plant heights were measured at four-day intervals starting at 0 dpi. Mock height attained (%) was calculated using the formula: $(H_I/H_m)*100$ where H_I is height of plant inoculated with treatment and H_m is height of mock-inoculated plant (Lapidot et al., 2006). DNA extracted at 28 dpi from plants infected with V30ΔC6-d was used as a template for PCR (C6 sequencing primers) with Sanger sequencing to confirm that mutation reversion to wild type sequence had not occurred. Plant trials were terminated after 36 dpi. Symptom scoring was done for three biological and 15 technical replicates. Student's t-tests were conducted using IBM SPSS Statistics v 28.0 software package to determine statistically significant differences between corresponding symptom scores in different treatments.

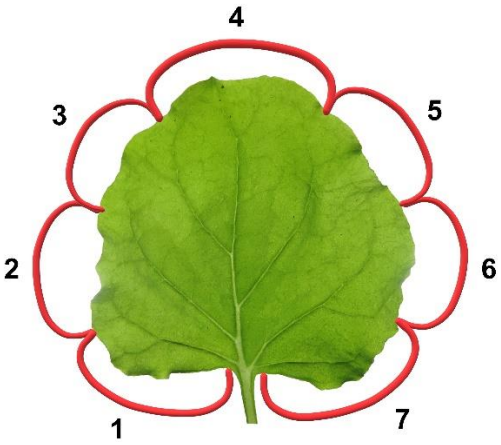


Fig. 2.1 *Nicotiana benthamiana* leaf edge zone demarcations used in upward leaf roll (ULR) symptom severity scoring system. Seven leaf zones are numbered.

Table 2.1 *Nicotiana benthamiana* upward leaf roll (ULR) symptom severity scoring system.

Score	Description
0	Leaf edges normal
1	Leaf edges turned up 1-7 zones, OR mild ULR 1-4 zones
2	Leaf edges turned up 2-7 zones AND mild/narrow ULR ≤4 zones, OR narrow/mild ULR 5-7 zones, OR broad/severe ULR 1-4 zones
3	Severe ULR ≥ 5 zones, OR entire leaf edge rolled upward and onto itself

Table 2.2 *Nicotiana benthamiana* swollen leaf vein (SV) symptom severity scoring system.

Score	Description
0	Leaf veins normal
1	Slight SV to mild SV over entire leaf, OR moderate SV over less than half of leaf
2	Moderate SV more than half of leaf only, OR moderate SV more than half of leaf with some severe SV, OR severe SV less than half of leaf
3	Severe SV more than half of leaf, OR dense SV over entire leaf

2.2.6 Nucleic acid extraction

For DNA extraction, two leaf discs (d = 12.7 mm) were collected from expanded leaves below the meristem at 12, 20, 28 and 36 dpi. Leaf tissue was flash frozen in liquid nitrogen and homogenised with 3 mm tungsten carbide beads (Qiagen) using the Qiagen TissueLyser II system (Qiagen) at 20 Hz for 1 min twice. CTAB extraction (Doyle, 1991) was done with the following modifications: an additional chloroform-isoamyl alcohol extraction step, two ethanol wash steps (95 % first wash, 75 % second wash), and incubation of samples at -80 °C for 30 min after the addition of isopropanol.

For total RNA extraction, ~100 mg of homogenized leaf material collected at 20 dpi was flash frozen in liquid nitrogen and homogenized with DEPC-treated 3 mm tungsten carbide beads (Qiagen) using the Qiagen TissueLyser II system (Qiagen) at 20 Hz for 1 min twice. RNA was extracted using TRIzol™ reagent (Thermo Fisher Scientific) as per the manufacturer's protocol with an additional chloroform separation step. All nucleic acids were quantified using the NanoDrop™ One spectrophotometer (Thermo Fisher Scientific) and purity was also assessed using gel electrophoresis. Extracted DNA and RNA was stored at -80 °C for downstream analysis.

2.2.7 qPCR

To determine viral load in inoculated plants at specified time points, real-time quantitative PCR (qPCR) was done using the Maxima SYBR Green/ROX qPCR Master Mix (Thermo Fisher Scientific) as per manufacturer's protocol with the CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories). Viral load was determined relative to the internal control glyceraldehyde 3-phosphate dehydrogenase (GAPDH) (Farooq et al., 2019). The degenerate primer pair used to target viral V2 ORF was designed using Primer3 with appropriate qPCR primer parameters selected (Kõressaar et al., 2018). The primer sequences are shown in S2 Table 2.2. Primer specificity was confirmed using standard PCR. Primer amplification efficiencies were determined through construction of standard curves (S2 Fig. 2.1). using CFX Maestro Software v 1.0 (Bio-Rad Laboratories). For qPCR experiments, three biological and nine technical replicates were used. Log viral fold change was calculated using the delta Ct method (Log $\Delta Ct=2^{-\Delta Ct}$). Student's t-tests were conducted using IBM SPSS

Statistics v 28.0 software package to determine statistically significant differences between viral load in different treatments.

2.2.8 RT-PCR

To investigate the presence of RNA transcripts derived from selected V30 and V22 genome regions, gene specific primers for cDNA synthesis were designed using Primer3 (Kõressaar et al., 2018). RT-PCR targets are described in Table 2.3. Primer sequences were evaluated for the presence of ss using the Oligoanalyzer™ Tool (https://www.idtdna.com/oligoanalyzer). For details of primers used refer to S2 Table 2.2. First strand cDNA was synthesized using the RevertAid™ First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) following the manufacturer's protocol using gene-specific primers with inclusion of all prescribed negative controls and mock-inoculated control. All cDNA was stored at -80 °C until subsequent PCR analysis was done using DreamTaq™ PCR Master Mix (Thermo Fisher Scientific). PCR was done in accordance with the manufacturer's instructions. PCR annealing temperatures, extension times and product sizes for all RT-PCR targets are shown in S2 Table 2.4. All RT-PCR products were resolved on 2 % agarose. RT-PCR was repeated three times.

Table 2.3 Virus region-specific RT-PCR products targeted in this study

RT-PCR product	**Target nt position (5'-3')[a]**	**Target description**
V30 IR-vs	2722-92	V30 IR spanning origin of replication
V30 IR-cs	92-2722	V30 IR spanning origin of replication
V30 V2-vs	56-249	V30 region spanning putative V2 ORF 5' extension (overlapping with IR 3' end) to canonical V2 ORF 5' region
V22 V2-vs	60-188	V22 region spanning IR/V2 ORF junction
V30 C6-cs	525-176	V30 putative C6 ORF 3' region
V30 R1-vs	142-635	V30 region 1 spanning V1/V2 ORF overlap
V30 R1-cs	635-142	V30 region 1 containing putative C5, C6, and C7 ORFs
V30 R2-vs	1231-1734	V30 region 2 containing putative V3, and V4 ORFs
V30 R2-cs	1734-1231	V30 region 2 spanning C1/C2 ORF and C2/C3 ORF overlaps

V30 R3-vs	367-1609	V30 region 3 spanning V2 ORF 3' region to putative V3 ORF
V30 R3-cs	1609-367	V30 region 3 spanning C1 ORF 3' region to putative C6 ORF

Key: [a], position on virus DNA-A monomer; vs, virion-sense; cs, complementary-sense.

2.3. Results

2.3.1 Investigation in silico of two variant genomes of ToCSV and other African begomoviruses reveals their variable coding potential.

Phylogenetic analysis of the DNA-A components of two ToCSV variant sequences [ZA:Mks30:08] and [ZA:Mks22:07] sharing 96 % sequence identity (Fig. 2.2A) revealed a distinct phylogenetic cluster with the exemplar isolate of ToCSV (AF261885). The closest relative of this ToCSV cluster was determined to be tomato leaf curl Diana virus (ToLCDiV: AM701765) isolated from Madagascar. Although [ZA:Mks30:08] and [ZA:Mks22:07] shared highest sequence identity with a Zimbabwean isolate of TbLCZV (AF350330) (84 % and 85 %), they were not closely related to this tobacco-infecting begomovirus. Pairwise nt sequence comparison of [ZA:Mks30:08] and [ZA:Mks22:07] with two ToCSV strain groups (ToCSV-Chok I and ToCSV-Chok II) identified in Mozambique (Sande et al., 2021) indicated that [ZA:Mks30:08] shared >96 % sequence identity with all ToCSV-Chok I isolates whilst [ZA:Mks22:07] shared <94 % sequence identity with isolates from both strain groups (data not shown).

The DNA-A monomers of agroinfectious clone constructs of the two ToCSV variant sequences, hereby termed V30 and V22 respectively, were subjected to ORF finder analysis (standard code, proteins ≥40 aa). The ORFs were mapped to the circular monomeric genomes (Fig. 2.2B). Both variant sequences were predicted to contain the six canonical (accepted) begomovirus ORFS (V1, V2, C1, C2, C3, C4). Two putative ORFs reported previously in other begomoviruses (C5 and C6) were identified. C5 was in frame with V1 and C6 was in frame with V2 but both putative ORFs were located on the cs strand. Three additional short putative ORFs with predicted protein sizes < 50 aa were also found. A short cs ORF with a start codon located downstream of C6 was termed C7. Two short vs ORFs were identified and provisionally named V3 and V4 based on the relative positions of their start codons.

The V2 ORF of V30 was predicted to contain a putative 5' extension (encoding 30 aa) upstream of the canonical V2 start codon (nt 139 to 141), extending into the 3' end of the IR. Both variant C1 ORFs were also found to contain putative 5' extensions (each encoding 12 aa) upstream of the canonical C1 start codon (nt 2612 to 2610 in V30, nt 2611 to 2609 in V22), extending into the 5' end of the IR. These putative 5' ORF extensions into the IR were absent in the majority of other African begomovirus sequences analysed. The nt positions of start/stop codons and predicted protein lengths for all accepted and putative V30 and V22 ORFs are shown in S2 Table 2.5.

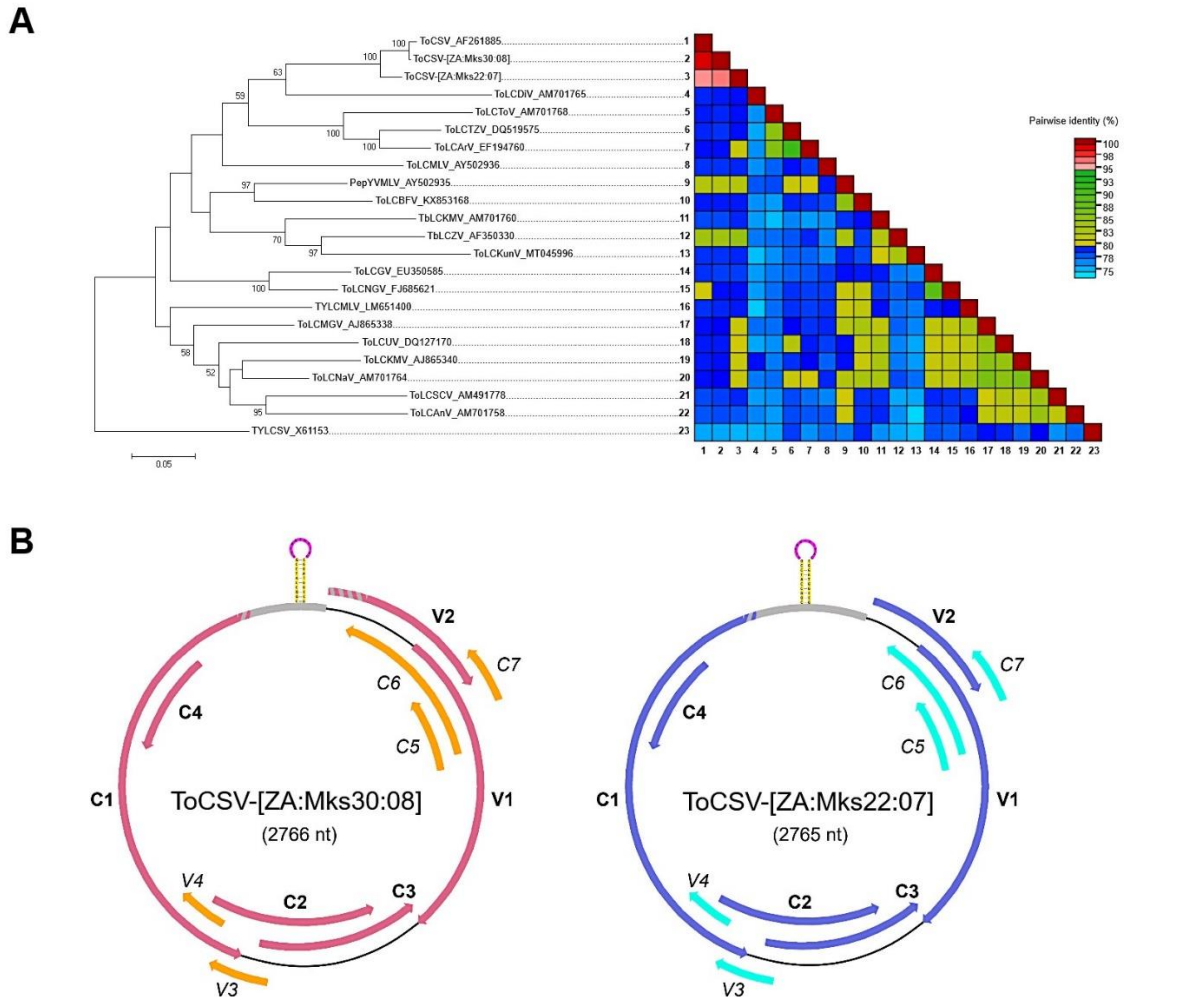


Fig. 2.2 Phylogenetic analysis of ToCSV variant sequences. **A**, Maximum likelihood phylogenetic tree constructed using full length DNA-A sequences of ToCSV-[ZA:Mks30:08] (V30), ToCSV-[ZA:Mks:22:07] (V22), and selected African monopartite tomato-infecting begomoviruses sharing ≥78 % DNA-A sequence identity with ToCSV (1,000 replications). Bootstrap values >50 % are shown. TYLCSV served as the outgroup. Corresponding three-colour pairwise sequence identity matrix with colour key indicating identities. A cut-off at 94 % served as the strain demarcation threshold for begomoviruses. **B**, Genome organisation of ToCSV-[ZA:Mks30:08] (V30), ToCSV-[ZA:Mks:22:07]

(V22) DNA-A genomes. Accepted virion-sense (V) and complementary-sense (C) ORFs coloured dark with bold labels, putative ORFs coloured light with italicised labels. Intergenic region (IR) in grey with stem-loop element shown. Putative 5' extensions of C1/V2 ORFs indicated using diagonal stripes. Details of all isolate sequences used in S2 Table 2.3.

Pairwise V30 and V22 protein aa sequence conservation and predicted ss alignments for six canonical (S2 Fig. 2.2 to 2.7) and five putative (S2 Fig. 2.8 to 2.12) proteins revealed several aa differences between the two variants, particularly in the V2 protein, TrAP (encoded by C2) and REn (encoded by C3). The V2 of V22 was shown to contain three additional C-terminal residues absent in V30 V2. The C6 protein of V30 contained a 35 aa C-terminal fragment absent in V22 C6. Predicted ss revealed that the CP (encoded by V1) contained predominantly strand elements. V2, V4, C5, and C7 proteins were predicted to contain helices but no strands were predicted.

The aa sequences of all identified canonical proteins (CP, V2, Rep, TrAP, REn, C4), canonical V2 and Rep with N-terminal extensions ($V2^N$ and Rep^N), and putative proteins (V3, V4, C5, C6, C7) were subjected to web-based software to predict unique features which could be used to infer their potential biological function (Table 2.4).

Table 2.4 Predicted protein features for ToCSV V30 and V22 proteins ≥40 aa in length

Feature	**Proteins**	
	V30	**V22**
NLS	CP, Rep, Rep^N	CP, Rep, Rep^N
cTP	none	none
mTP	V4	V4
SP	$V2^N$, V3, C6, C7	V3, C6, C7
TMH	$V2^N$, C6, C7	C6

Key: ^{N}, protein including putative N-terminal extension; NLS, nuclear localisation signal; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide; SP, signal peptide; TMH, transmembrane helix.

The CP and Rep proteins of V30 and V22 were predicted to contain NLS but no other canonical protein was predicted to contain any features of interest. No proteins were predicted to contain cTP. Putative V3, C6 and C7 proteins were predicted to contain SP, whilst V4 was predicted to contain mTP. Interestingly, SP and TMH were predicted

to exist in the N-terminal region of V30 $V2^N$ which was absent in V30 V2. No cTP were identified in any of the proteins.

Once the ORFs present in V30 and V22 were identified, the ToCSV exemplar isolate (AF261885) and selected African mono- and bipartite begomovirus isolates sharing ≥ 75% DNA-A sequence identity with ToCSV were screened manually for putative ORFs of interest using ORF finder (standard code, proteins ≥40 aa). African begomovirus Isolates shown to contain truncated/fragmented V1, V2, or C1 ORFs were excluded. Of the 60 accepted isolates screened: 15 (25 %) had V3/AV3, 24 (40 %) had V4/AV4, 53 (88 %) had C5/AC5, 39 (65 %) had C6/AC6, and 28 (47 %) had C7.

Phylogenetic analysis of full length V3/AV3 proteins (S2 Fig. 2.13) showed three highly divergent clusters, separated according to virus host plant. Multiple aa sequence conservation and ss alignments (S2 Fig. 2.14 to 2.16) confirmed the variability of the putative V3/AV3 protein. Phylogenetic analysis of putative V4/AV4 proteins also showed divergent aa sequences (S2 Fig. 2.17), although the length of most putative V4/AV4 proteins was identical (42 aa). Multiple V4/AV4 aa sequence and ss alignments revealed a highly conserved C-terminal region predicted to contain two consensus helices (S2 Fig. 2.18). Although the putative C7 ORF was found in monopartite begomoviruses only, the predicted proteins were also found to be highly divergent (S2 Fig. 2.19) with predicted ss varying considerably (S2 Fig. 2.20).

2.3.2 Two variant agroinfectious clones of ToCSV induce significantly differing pathologies in *N. benthamiana*.

Infectivity assays using ToCSV V30 and V22 agroinfectious clones were performed for the first time in *N. benthamiana*. At ~12 dpi, virus-like symptoms began to appear in the emerging leaves of plants separately agroinoculated with V30 and V22. Leaf epinasty, rugosity, mild chlorosis, and leaf curling were observed (Fig. 2.3A). Mock-inoculated plants appeared identical to non-inoculated plants and developed normally with no disease symptoms being observed (S2 Fig. 2.21 and S2 Fig. 2.22). Two distinct symptom types, ULR and SV, were consistently reported in the infected plants but differed in their severity between the two variants (S2 Fig. 2.23 and S2 Fig. 2.24). The ULR disease symptom initially presented as upward turning of the edges of the emerging leaf, gradually progressing to upward rolling of the leaf edges as the leaf

expanded. Increased severity of ULR was associated with leaf roll being present along the entire leaf edge or broad leaf roll that resulted in the upward rolling of the entire leaf surface. The SV phenotype manifested as abnormal swelling of the leaf veins resulting in abaxial vein bulging and adaxial vein depression. This symptom was especially apparent when viewed through a light source as an enhanced visual contrast of the vein network was seen. Severe SV symptoms presented as a dense network of affected leaf veins often associated with rugosity and leaf distortion. Milder SV presented as sparsely distributed SV or slight SV present near the leaf edges without associated rugosity.

Disease progression was markedly different for the two variants. In V22-infected plants, moderate to severe ULR was evident from 20 to 24 dpi and declined thereafter to near zero at 36 dpi. In V30, ULR was mild to almost absent over the entire monitoring period (Fig. 2.3B). A moderate to severe SV symptom phenotype was observed in both variants, peaking between 20 and 24 dpi. A significant recovery in SV was seen in V22-infected plants between 24 and 36 dpi, unlike in V30-infected plants which did not show decreased SV severity by 36 dpi (Fig. 2.3B). Plant stunting was observed in both variants with V30-infected plants displaying more severe stunting than V22 at 36 dpi (S2 Fig. 2.21). Infection was confirmed using conserved V2 ORF primer-PCR and Sanger sequencing (data not shown).

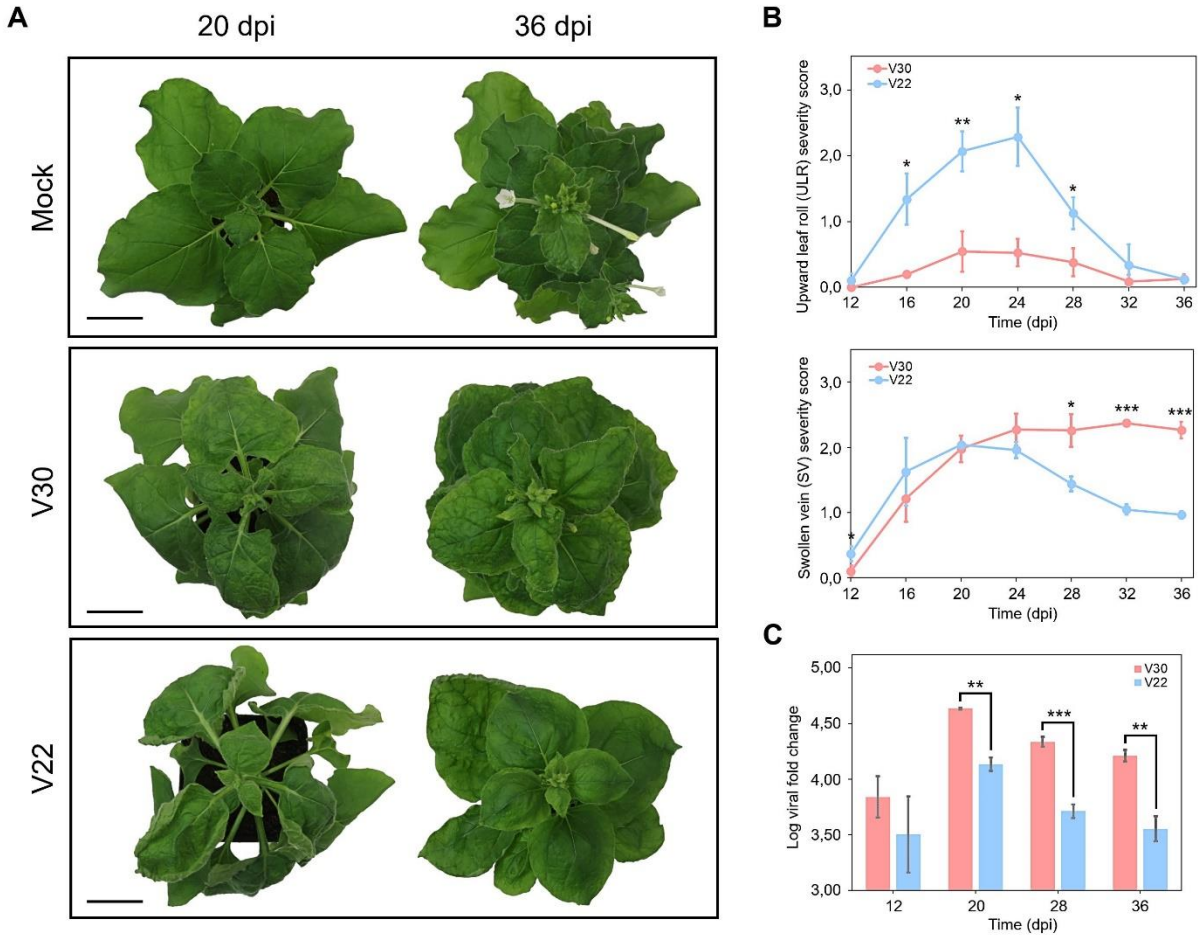


Fig. 2.3 Infection of *Nicotiana benthamiana* with ToCSV wild type infectious clones in single infection. **A,** Symptoms in plants inoculated with mock (*Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300), V30, and V22 at 20 dpi (left) and 36 dpi (right). Bar = 2 cm. **B,** Corresponding upward leaf roll (ULR) and swollen vein (SV) symptom severity scores at four-day intervals from 12 to 36 dpi. **C,** Log viral fold change at 20 and 36 dpi. Bars represent mean ± SD. Student's t-test levels of significance: *, P < 0.05; **, P < 0.01; ***, P < 0.001.

Viral load relative to the internal control GAPDH was determined using qPCR (Fig. 2.3C). Plants infected with V30 had significantly higher relative viral load than V22 at 20 dpi ($P < 0.01$), 28 dpi ($P < 0.001$) and 36 dpi ($P < 0.01$). Peak viral titre in both variant-infected plants was observed at 20 dpi. To investigate the presence of long genome-derived transcript RNAs in V30-infected plants, RT-PCR using virus-specific primers was done. Three RNA transcript targets were mapped to canonical and putative ORF coding regions of both vs and cs strands (Fig. 2.4). Viral RNA transcripts were predicted to span multiple ORFs, and none were entirely derived from a region spanning a single ORF. All target RT-PCR products were identified in V30-infected plants, but not in mock-inoculated plants (Fig. 2.4).

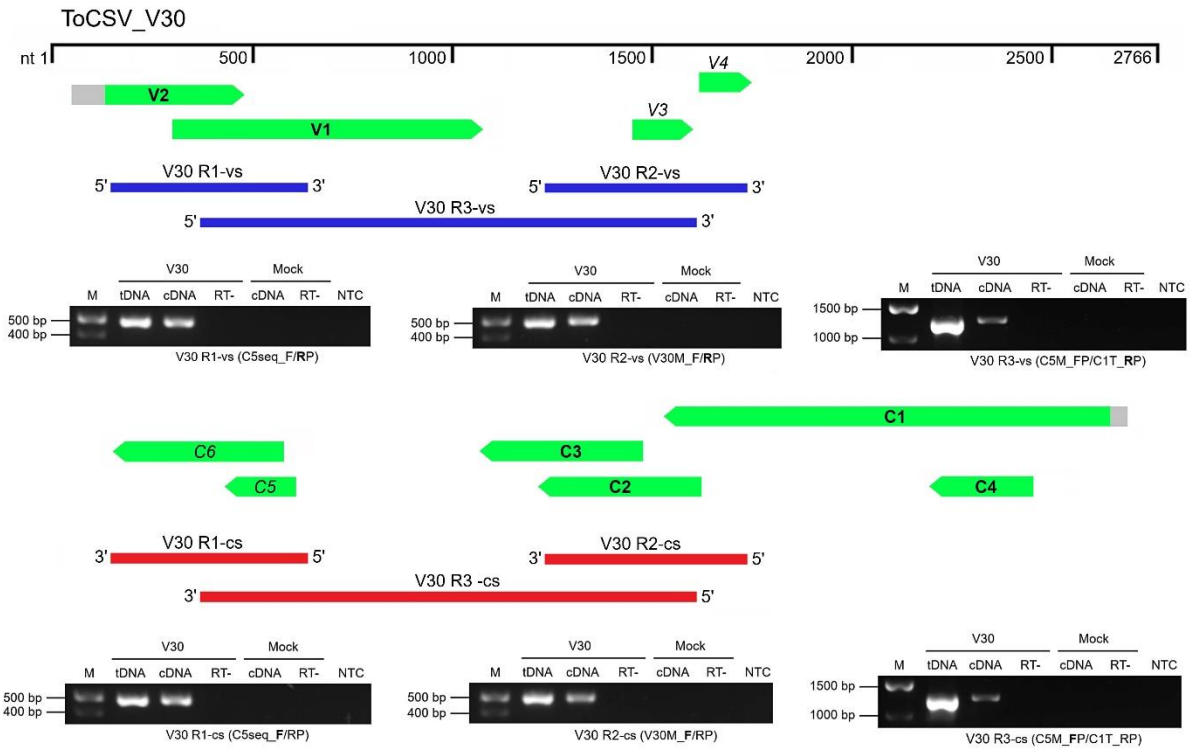


Fig. 2.4 Identification of ToCSV V30 RNA transcripts derived from three regions (R1, R2, R3) using RT-PCR. Accepted (bold) and putative (italicised) ORFs segregated by virion (vs) and complementary sense (cs) frames. Positions of three vs (blue) and cs (red) strand-derived RT-PCR products relative to ORF positions indicated. Agarose gel images of three vs (top) and three cs (bottom) RT-PCR products. M, DNA ladder marker; tDNA, PCR of total DNA extracted from plants inoculated with V30 (positive control); cDNA, RT-PCR of total RNA extracted from plants inoculated with V30 or mock; RT-, negative reverse transcriptase enzyme control; NTC, PCR no template control. Primers used for PCR in brackets with primer used for cDNA synthesis highlighted in bold. Mock is RT-PCR of total RNA extracted from plants inoculated with *Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300.

2.3.3 The 3' IR of ToCSV acts as a symptom phenotype determinant in *N. benthamiana*.

Pairwise sequence alignment of the full-length IR of both wild type ToCSV variant sequences (S2 Fig. 2.25) revealed the presence of two short (~30 nt) regions in the 3' IR having sequence differences. These short regions were located to the left and right of a conserved TATA box-containing region (nt 95-107) present upstream of the canonical V2 start codon. A multiple nt sequence alignment of the 3' IR of several isolates of ToCSV and related African tomato-infecting begomoviruses was done to identify recognised sequence motifs and regulatory elements present in the 3' end (Fig. 2.5). Elements identified included the conserved stem-loop element containing

nonanucleotide, CLE-like motif, TATA box, and the TACE. The stem-loop element was highly conserved amongst begomovirus isolates. All isolates of ToCSV and most other begomovirus isolates were shown to contain a single CLE-like motif.

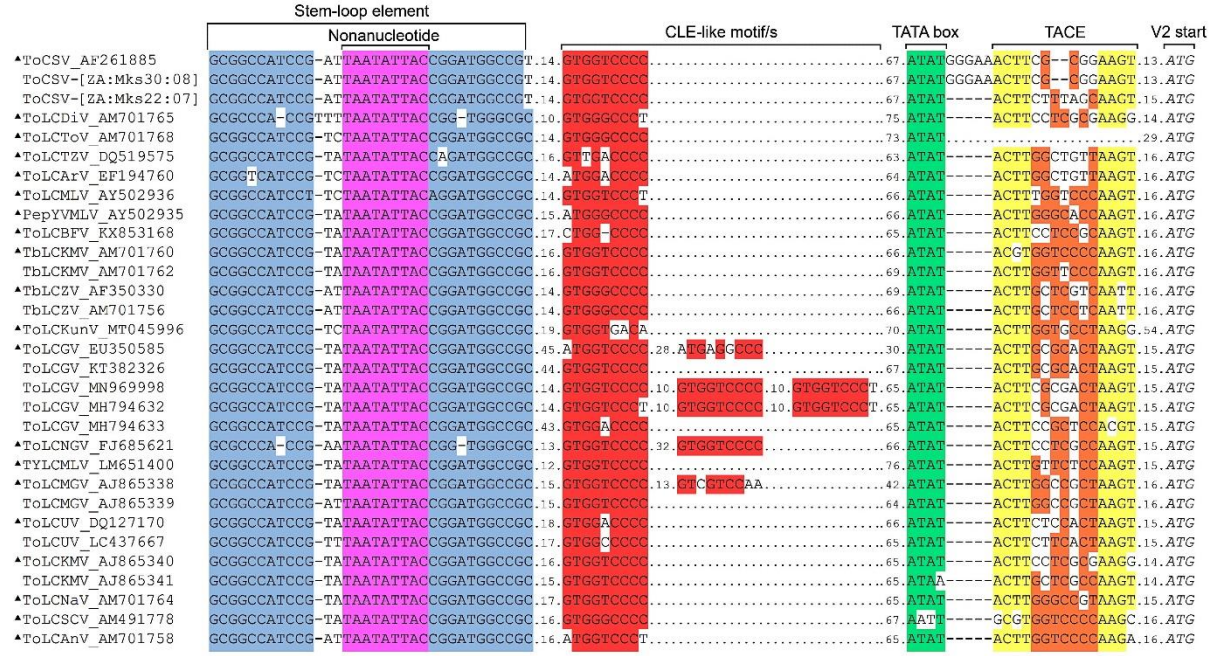


Fig. 2.5 Partial intergenic region (IR) viral motif/regulatory element sequence alignment. IR region spanning stem-loop element to V2 start codon (italicised) indicated for ToCSV-[ZA:Mks30:08], ToCSV-[ZA:Mks:22:07], and selected African monopartite tomato-infecting begomoviruses sharing ≥78 % DNA-A sequence identity with ToCSV. Exemplar isolates indicated with black triangles. CLE and TACE refer to conserved late element and TATA-associated composite element, respectively. Highlighted colours indicate motif consensus sequence present. Hyphens indicate alignment-generated gaps, ellipsis indicates spacer sequence with nt number shown. Details of all isolate sequences used in S2 Table 2.3.

All begomovirus isolates contained a TACE downstream of the TATA box, except for the exemplar isolate of tomato leaf curl Toliara virus (ToLCToV: AM701768). The central spacer region of the TACE between the conserved left (ACTT) and right (AAGT) arms was highly variable. Multiple sequence comparison uncovered two features of the TACE found only in the ToCSV exemplar isolate (AF261885) and ToCSV-[ZA:Mks30:08]. The first feature was an additional spacer sequence (GGGAA) between the TATA box and the conserved left arm sequence of the TACE. The second feature was a reduced TACE central spacer region. This spacer region contained five

nt instead of the consensus seven nt present in the TACE of ToCSV-[ZA:Mks:22:07] and other African tomato-infecting begomoviruses.

Three partial V30 IR 3' sequence swap mutants (Fig. 2.6A) were used to elucidate the potential role that these IR sequence differences may play in pathogenicity in *N. benthamiana*. Refer to S2 Fig. 2.26 for multiple sequence alignment of the 3' IR of V30 and the three IR swap mutants. Agroinoculation of *N. benthamiana* with V30 IR mutants resulted in significant changes to the wild type V30 symptom phenotype. Severe and persistent ULR symptoms were observed to develop in plants inoculated with V30ΔIR-s (Fig. 2.6B, C; S2 Fig. 2.27). The severity in SV symptoms was also significantly increased between 24 and 36 dpi when compared to V30-infected plants. A dense network of SV was associated with severe rugosity. Both ULR and SV persisted without recovery by 36 dpi. Symptoms induced by V30ΔIR-sl were identical to those observed in plants inoculated with V30 (Fig. 2.6B, C; S2 Fig. 2.28). Plants infected with V30ΔIR-sr developed severe ULR and SV with severity scores being significantly higher than in V30-infected plants, although mean ULR scores were lower than was recorded in V30ΔIR-s-infected plants (Fig. 2.6B, C; S2 Fig. 2.29). Measurement of viral load relative to GAPDH (Fig. 2.6D) showed a significantly higher viral titre at 28 dpi in plants infected with V30ΔIR-s than in V30-infected plants ($P < 0.01$). To probe for IR-derived RNA transcripts crossing the Ori in V30-infected plants, RT-PCR using IR-specific primers targeting vs and cs RNA transcripts was done. RT-PCR products corresponding to RNAs derived from both strands were positively identified in RNA extracted from V30-infected plants but not mock-inoculated plants (Fig. 2.6E).

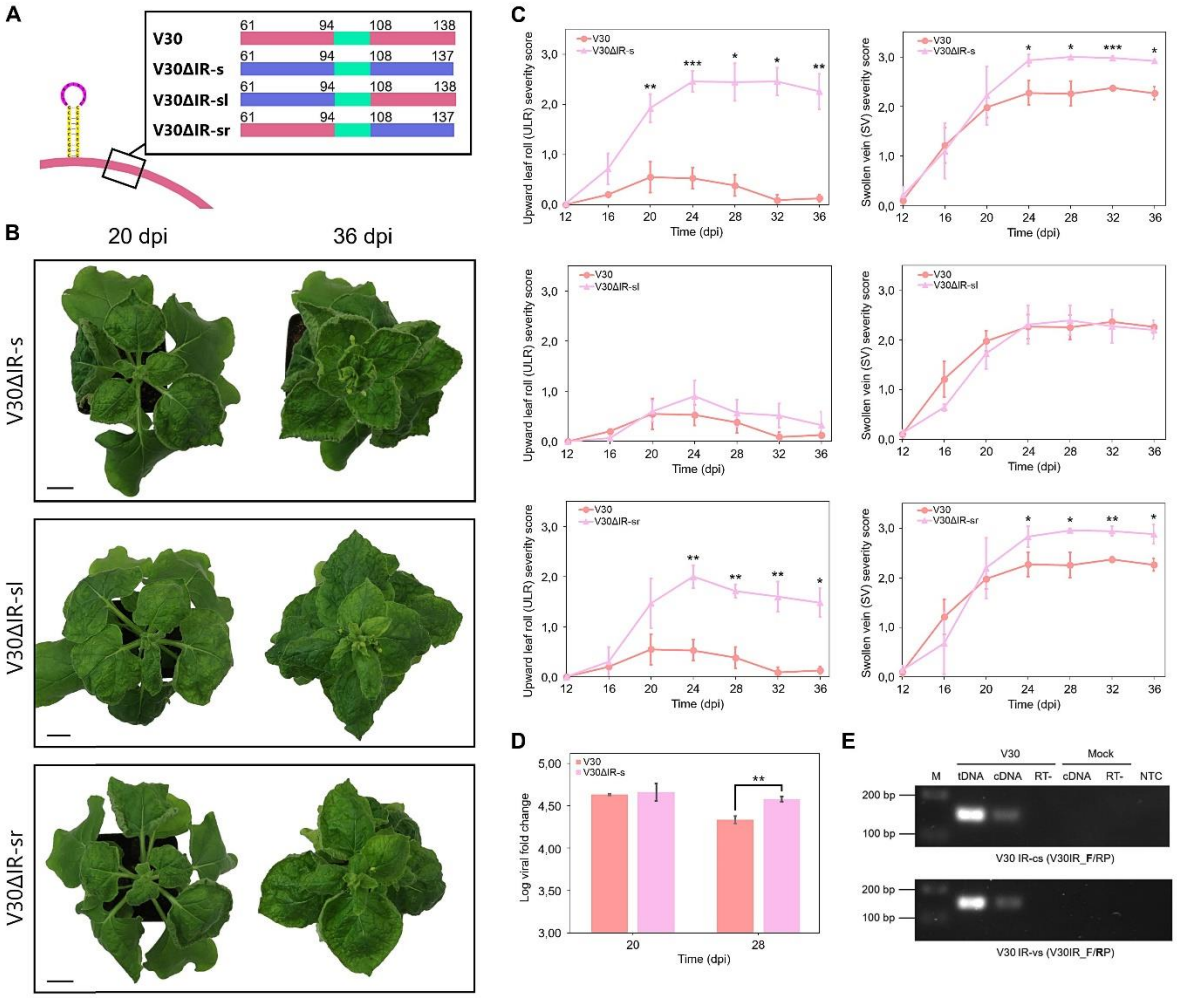


Fig. 2.6 Infection of *Nicotiana benthamiana* with ToCSV V30 IR-swap mutant infectious clones in single infection. **A**, Summary of IR-swap sequence identities: V30 sequence in red, V22 sequence in blue with boundaries shown as nt position numbers. Conserved region containing TATA box in green. **B**, Symptoms in plants inoculated with V30ΔIR-s, V30ΔIR-sl, and V30ΔIR-sr at 20 dpi (left) and 36 dpi (right). Bar = 2 cm. **C**, Corresponding upward leaf roll (ULR) and swollen vein (SV) symptom severity scores at four-day intervals from 12 to 36 dpi. **D**, Log viral fold change at 20 and 28 dpi. **E**, Agarose gel images of V30 complementary (cs) and virion sense (vs) IR-derived RT-PCR products. M, DNA ladder marker; tDNA, PCR of total DNA extracted from plants inoculated with V30 (positive control); cDNA, RT-PCR of total RNA extracted from plants inoculated with V30 or mock; RT-, negative reverse transcriptase enzyme control; NTC, PCR no template control. Primers used for PCR in brackets with primer used for cDNA synthesis highlighted in bold. Mock is RT-PCR of total RNA extracted from plants inoculated with *Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300. Bars represent mean ± SD. Student's t-test levels of significance: *, P < 0.05; **, P < 0.01; ***, P < 0.001.

It is important to consider that the start codon of the putative V30 V2 N-terminal segment is present in the 3' IR of V30. This feature exists because the 3' IR of V30

contains one extra nt in comparison to the 3' IR of V22. This consequently places the putative V2 start codon in-frame with the canonical V2 start (S2 Fig. 2.25). The putative V2 start (nt 49 to 51) is located upstream of the IR 3' fragment of interest. Due to IR swap mutation, the V2 putative N-terminal segment was predicted to be lost in the V2 protein encoded by both V30ΔIR-s and V30ΔIR-sr. Conversely, the V2 protein of V30ΔIR-sl was predicted to retain this segment although its sequence identity differed from that of V30 (S2 Fig. 2.30A, B). Analysis of predicted protein features showed that both N-terminal segments were predicted to contain helix ss elements (S2 Fig. 2.30C), as well as TMH being predicted by at least one web-based prediction server (S2 Fig. 2.30D).

2.3.4 The V2 ORF of ToCSV is implicated in modulating disease severity and subsequent symptom recovery in *N. benthamiana.*

In addition to the IR, the coding region spanning the canonical V2 ORF of both ToCSV variants also contains nt sequence differences. The two canonical V30 and V22 V2 protein sequences share 94 % aa sequence identity. Closer examination of the two variant V2 proteins and selected African tomato-infecting begomovirus V2 proteins using phylogenetic analysis revealed that ToCSV V2 proteins formed a distinct cluster (Fig. 2.7A). The V30 V2 protein sequence was identical to the V2 of the ToCSV exemplar isolate (AF261885). Multiple aa sequence alignment of selected V2 proteins (S2 Fig. 2.31) showed that the V2 protein contained regions of high sequence conservation and was predicted to contain five consensus helix ss elements. A V2 region corresponding to the first three predicted helices (aa 13 to 61) was mostly conserved between isolate sequences (Fig. 2.7B). V30 and V22 V2 pairwise aa sequence comparison (S2 Fig. 2.3) revealed a total of seven aa differences present in the first 58 aa of the two V2 proteins. The African tomato-infecting begomovirus V2 multiple sequence alignment was used to construct a sequence logo of this region of interest (Fig. 2.7C). The first helix (19 aa) contained a series of highly conserved residues, with a consensus sequence KYLQ followed by a short region of higher variability in the C-terminus. The second helix (14 aa) contained three conserved leucine residues (hydrophobic), one conserved aspartic acid (negatively charged), and a highly variable residue at aa position 11. The helix terminated in a conserved

arginine (positively charged). The third and shortest helix contained a consensus sequence YVEA(T/S) with tyrosine being hydrophobic, glutamic acid being negatively charged and threonine/serine being polar and uncharged. Hydrophobic regions were mapped to the first two helices (Fig. 2.7C). The consensus sequence of the putative PKC phosphorylation motif was TRR and the threonine of this motif was located at the C-terminal end of the third helix (Fig. 2.7C). Interestingly, the PKC motif sequence present in the V2 proteins of V30 and the exemplar isolate was SRR (Fig. 2.7B).

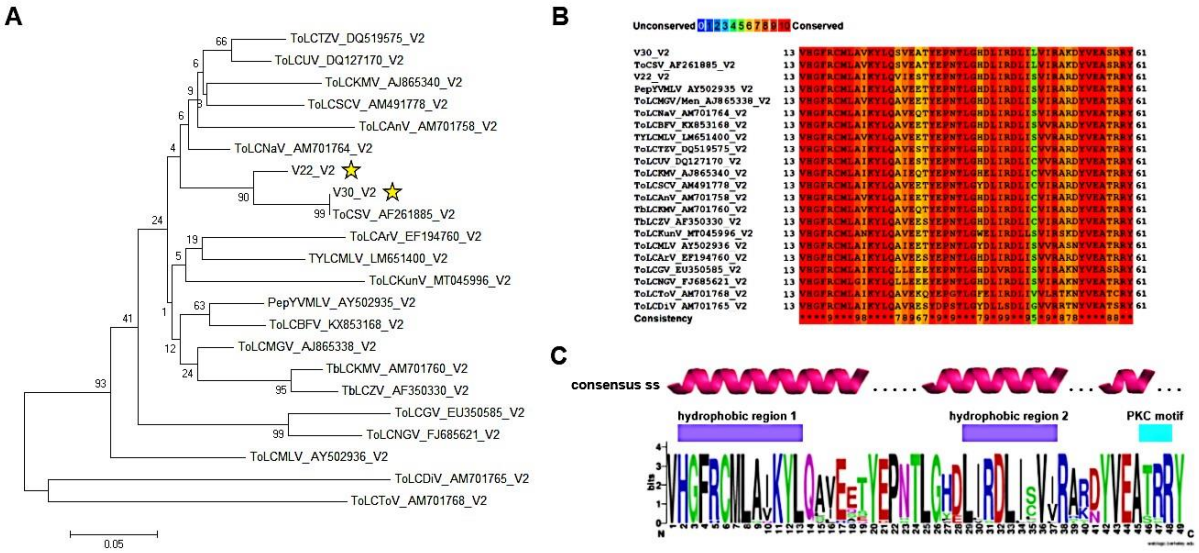


Fig. 2.7 Phylogenetic and aa sequence analysis of begomovirus V2 proteins. **A**, Phylogenetic tree constructed using full length accepted V2 aa sequences of ToCSV V30, V22, and selected African monopartite tomato-infecting begomoviruses sharing ≥78 % DNA-A sequence identity with ToCSV. Alignment done using the neighbour-joining method (1,000 replications). V30 and V22 indicated using stars. **B**, Partial V2 aa multiple sequence alignment with colour key indicating aa conservation. Alignment generated using PRALINE. **C,** Partial V2 consensus aa sequence logo with positions of PSIPRED-predicted consensus helices (magenta cartoon), two hydrophobic regions (violet boxes) and putative protein kinase C (PKC) phosphorylation motif (cyan box) shown. Full length PRALINE aa alignment sequence conservation and PSIPRED predicted ss elements with partial alignment sequence indicated with black box in S2 Fig. 2.31. Details of all isolate sequences used in S2 Table 2.3.

To examine the role that the V2 coding region may exert in the pathogenicity of the two variants, full length V2 ORF swap mutants were synthesised: V30ΔV2-s (V30 V2 swap mutant) and V22ΔV2-s (V22 V2 swap mutant). The V1, C5, C6 and C7 ORFs overlap with V2 and thus swapping of this coding region resulted in aa changes to the proteins encoded by the overlapping ORFs (S2 Table 2.6). Genome organisation and nt swap coordinates for both V2 swap mutants is shown in Fig. 2.8A. Only the

canonical V2 coding region was swapped and as a result, the putative V30 N-terminal segment was encoded by V30ΔV2-s, but not by V22ΔV2-s. Plants infected with V30ΔV2-s displayed symptoms nearly identical to V30, with no increase in ULR symptom severity. A decrease in mean SV severity was observed between 24 and 36 dpi (Fig. 2.8B, C; S2 Fig. 2.32). At 36 dpi, the mean SV severity score was significantly lower than that recorded for V30-infected plants. This apparent recovery in SV symptoms was associated with a substantial reduction in leaf surface rugosity when compared to plants infected with V30 (S2 Fig. 2.32).

In contrast, plants infected with V22ΔV2-s displayed a significant increase in both severity and duration of ULR and SV with a loss of recovery phenotype (Fig. 2.8B, C). The broad upward rolling of the entire leaf edge resulted in severe leaf distortion (S2 Fig. 2.33). After 28 dpi mean symptom severity decreased but by 36 dpi both ULR and SV remained at peak severity levels seen in V22-infected plants.

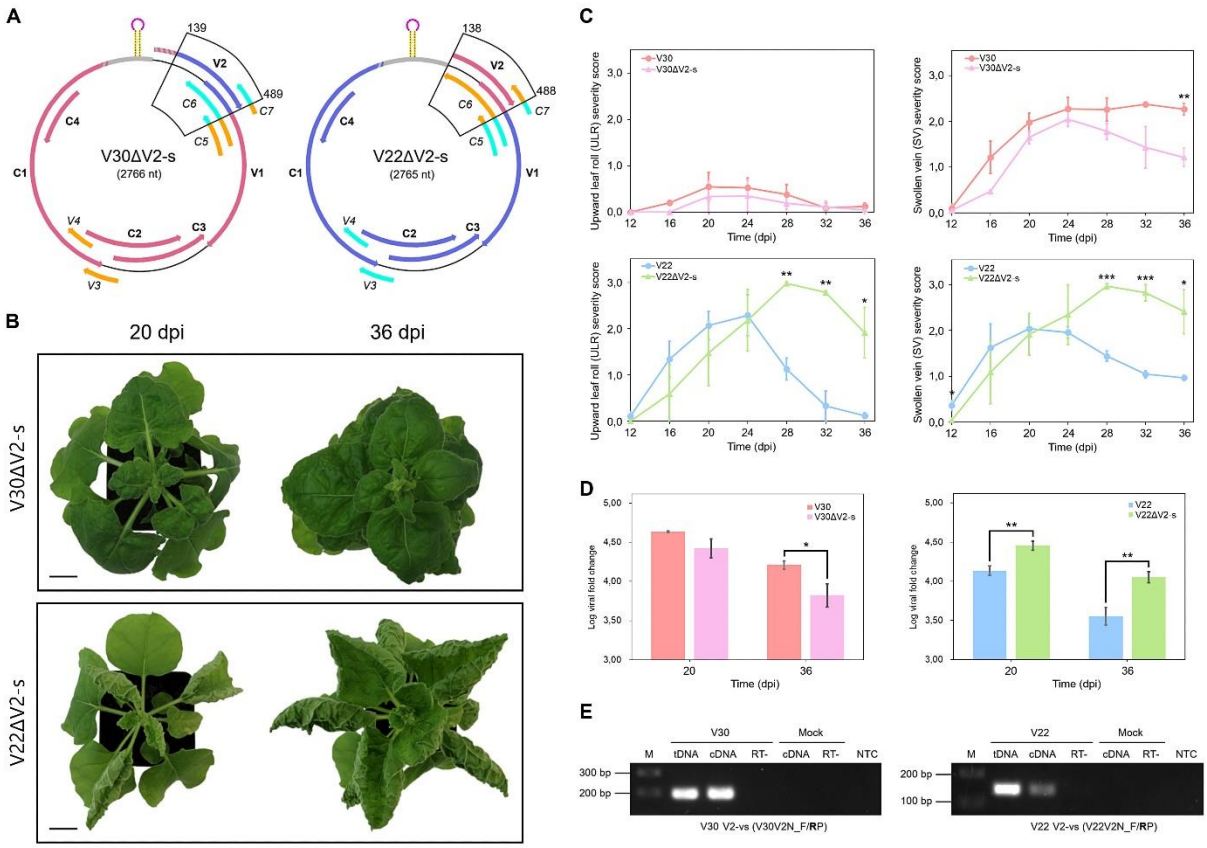


Fig. 2.8 Infection of Nicotiana benthamiana with ToCSV V30 and V22 V2-swap mutant infectious clones in single infection. **A**, Genome organisation of V30ΔV2-s and V22ΔV2-s DNA-A genome monomers. Accepted virion-sense (V) and complementary-sense (C) ORFs coloured dark with bold labels, putative ORFs coloured light with italicised labels. Intergenic region (IR) in grey with stem-loop element shown.

Putative 5' extensions of C1/V2 ORFs indicated using diagonal stripes. Colour-coded swapped fragments in fan-shaped boxes with nt boundary positions indicated. **B**, Symptoms in plants inoculated with V30ΔV2-s, and V22ΔV2-s at 20 dpi (left) and 36 dpi (right). Bar = 2 cm. **C**, Corresponding upward leaf roll (ULR) and swollen vein (SV) symptom severity scores at four-day intervals from 12-36 dpi. **D**, Log viral fold change at 20 and 36 dpi. **E**, Agarose gel images of V30 virion sense (vs) putative V2 ORF 5' extension region-derived (left) and V22 vs IR/V2 interface-derived (right) RT-PCR products. M, DNA ladder marker; tDNA, PCR of total DNA extracted from plants inoculated with V30 or V22 (positive control); cDNA, RT-PCR of total RNA extracted from plants inoculated with V30, V22, or mock; RT-, negative reverse transcriptase enzyme control; NTC, PCR no template control. Primers used for PCR in brackets with primer used for cDNA synthesis highlighted in bold. Mock is RT-PCR of total RNA extracted from plants inoculated with *Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300. Bars represent mean ± SD. Student's t-test levels of significance: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Relative viral load in V2 swap mutant-infected plants was measured and compared with that of the corresponding wild type variant-infected plants (Fig. 2.8D). Plants infected with V30ΔV2-s had significantly lower viral load than plants infected with V30 at 36 dpi ($P < 0.05$). A very significant difference in viral load was seen in V22ΔV2-s-infected plants at both 20 and 36 dpi compared to V22-infected plants ($P < 0.01$). Viral load was higher at these two timepoints compared to viral load measured in plants infected with V22 (Fig. 2.8D). The presence of a putative V30 V2 ORF 5' region vs transcript was determined using RT-PCR with V2-specific primers. The target RT-PCR product was positively identified in total RNA extracted from V30-infected but not mock-inoculated plants, indicating the presence of a transcript originating from the V2 putative 5' region and extending into the consensus V2 ORF (Fig. 2.8E). Although V22 is not predicted to contain the V2 ORF 5' extension, RT-PCR of RNA extracted from V22-infected plants was also conducted using V22-specific primers targeting the same region as was done for V30 and this yielded a positive result (Fig. 2.8E). This indicated the presence of an RNA transcript spanning the V2/IR junction in V22-infected plants in the absence of a predicted putative V2 ORF 5' extension.

2.3.5 C5/AC5 and C6/AC6 – additional players in the game that require further consideration.

The two putative C5/AC5 and C6/AC6 ORFs have gained considerable interest in recent studies relating to begomovirus pathogenicity. Previously neglected due to lack of sequence conservation and variability in length, the current study aimed to examine the prevalence and sequence characteristics of these two ORFs in selected African begomoviruses. Both C5/AC5 and C6/AC6 ORFs are located in the V2/V1 or AV2/AV1 coding region and are transcribed from the cs strand. The key defining feature which was used to differentiate these two ORFs was that C5/AC5 was in frame with V1/AV1 and C6/AC6 was in frame with V2/AV2. This relative ORF arrangement was consistent for all begomovirus isolates analysed in this study.

A total of 55 African begomovirus isolate sequences (including V30 and V22) were predicted to contain a C5/AC5 ORF. Analysis of putative C5/AC5 proteins proved to be complex due to two key factors. Protein length was shown to be highly variable, ranging from 42 to 262 aa. Segmented ORF profiles were evident with over half of isolates analysed having two distinct in-frame C5/AC5 ORF segments encoding two separate proteins. These shorter proteins were found to align with longer continuous C5/AC5 proteins encoded by other isolates. Multiple sequence alignments were conducted to segregate all C5/AC5 proteins identified by regions of aa conservation. The conserved domain database (CDD) recognises two C5/AC5 domains: pfam04807 and pfam08464 hereby termed “AC5-1” and “AC5-2”, respectively. Sequence analysis revealed a putative C5/AC5 protein N-terminal long helix region (NLHr) which was identified in 20 of the 55 C5/AC5-containing isolates. A summary of C5/AC5 domain organisation and translational profiles is shown in Fig. 2.9A, B. Phylogenetic analysis of C5/AC5 proteins segregated into the three domain groups (S2 Fig. 2.34 to 2.36) revealed that C5/AC5 protein sequences were highly divergent and ToCSV C5 proteins formed a distinct cluster. Both V30 and V22 C5 proteins contained the AC5-2 domain only. Multiple aa sequence alignment of the three segregated C5 protein groups, with domains of interest indicated with black boxes, can be seen in S2 Fig. 2.37 to 2.39.

The NLHr was so named due to the presence of a predicted consensus 26 aa-long helix ss structure. A portion of this long helix was predicted to exist as a

transmembrane region with a series of conserved hydrophobic residues being present (Fig. 2.9C; S2 Fig. 2.40). A conserved histidine (positively charged) was located at the N-terminal end of the long helix and all NLHr identified contained two to three glutamine residues (polar, uncharged) also at the N-terminal end. The AC5-1 domain was predicted to contain two strand ss elements, with the second strand followed immediately by a helix. Both stands each contained three consensus hydrophobic residues. A conserved histidine (positively charged) was situated at the junction between the second strand and the helix which contained two lysine residues (positively charged) in most of the AC5-1 domains identified (Fig. 2.9D; S2 Fig. 2.41). The AC5-2 domain was predicted to contain two helices. Four highly conserved leucine residues (hydrophobic), and two highly conserved threonine residues (polar, uncharged) were situated at the N-terminal end. Two highly conserved proline residues were located at the C-terminal end (Fig. 2.9E; S2 Fig. 2.42).

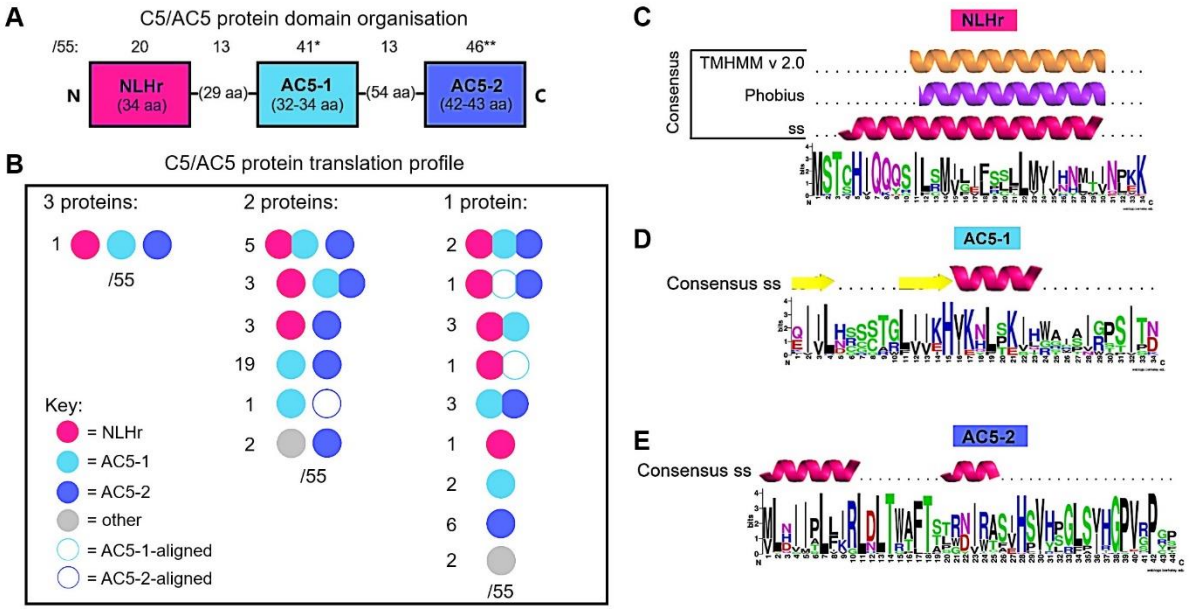


Fig. 2.9 C5/AC5 protein domain and aa sequence analysis for selected African begomoviruses. **A**, Domain organisation of predicted C5/AC5 proteins ≥ 40 aa in length following standard code in 55 African mono- and bipartite begomovirus DNA-A isolates sharing ≥75 % DNA-A sequence identity with ToCSV revealed three domain regions and two spacer regions. Two domains (AC5-1 and AC5-2) are recognised in the conserved domain database (CDD – pfam04807 and pfam08464 respectively). A novel putative domain was identified: NLHr, N-terminal long helix region. The length of each domain/spacer and the number of isolate sequences predicted to have C5/AC5 proteins containing these regions is indicated. *, 41/43 aligning isolate sequences contain a recognised AC5-1 domain. **, 46/47 aligning isolate sequences contain a recognised AC5-2 domain. **B**, Segmented translation profile of C5/AC5 ORF identified following standard code. C5/AC5 protein number and number of isolate

sequences predicted to have each profile are shown. Key indicates domain organisation present. **C**, NLHr multiple sequence alignment consensus: PSIPRED predicted consensus secondary structure (ss), predicted transmembrane helices (Phobius and TMHMM v 2.0), and aa sequence logo. **D**, AC5-1 multiple sequence alignment consensus: PSIPRED predicted consensus ss and aa sequence logo. **E**, AC5-2 multiple sequence alignment consensus: PSIPRED predicted consensus ss, and aa sequence logo. Helices shown as magenta cartoons, strands shown as yellow cartoons. Full length PRALINE aa alignment sequence conservation with consensus sequence indicated with black boxes shown in S2 Fig. 2.37 to 2.39. Partial PSIPRED predicted ss elements, transmembrane helix prediction and CDD domain identification in S2 Fig. 2.40 to 2.42. Details of all isolate sequences used in S2 Table 2.3.

The C6/AC6 ORF is more rarely found than C5/AC5 and consequently very little is known about the role of the putative C6/AC6 protein in begomovirus infection. The genomes of V30, V22 and the ToCSV exemplar isolate (AF261885) were predicted to contain putative C6 proteins >100 aa in length. Of the 39 additional African begomovirus isolates found to contain C6/AC6, only three cassava-infecting begomovirus putative AC6 proteins were >100 aa in length. Although the majority had one C6/AC6 protein, three begomovirus isolates were predicted to contain a discontinuous C6/AC6 ORF encoding two distinct protein segments (S2 Fig. 2.43). Phylogenetic analysis confirmed that C6/AC6 proteins are highly divergent, varying greatly in length and sequence identity (S2 Fig. 2.43 and 2.44). The C6 ORF of V30 encodes an additional 35 aa C-terminal segment which is absent in V22 C6 (S2 Fig. 2.11). Multiple sequence alignment of C6/AC6 proteins revealed that C6/AC6 proteins of 34 of the 39 isolates contain a 39 aa-long region herein termed the conserved core domain (C6cd) (S2 Fig. 2.44). The aa sequence of the C6cd is identical for the C6 proteins of V30 and V22 except for three non-conserved residues at the C-terminal end (S2 Fig. 2.11). The consensus ss of the C6cd was predicted to include an N-terminal helix and three short strand elements (Fig. 2.10). The helix and first strand both contained a conserved arginine (positively charged) with both elements being separated by a consensus sequence GLSKF. The second and third strands each contained a conserved isoleucine (hydrophobic) and were found to be separated by the consensus sequence S(P/A)GRF. A highly conserved aspartic acid (negatively charged) formed part of the third helix which terminated with a non-conserved residue.

Fig. 2.10 Sequence logo of C6/AC6 putative conserved core domain (C6cd) constructed using C6/AC6 aa sequences of ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C6/AC6 proteins containing C6cd. PSIPRED predicted consensus secondary structure (ss) indicated above sequence logo. Helix shown as magenta cartoon, strands shown as yellow cartoons. Phylogenetic tree of all C6/AC6 aa sequences used, C6cd PRALINE aa alignment sequence conservation and PSIPRED predicted ss elements in S2 Fig. 2.44.

To investigate the effect of removing the V30 C6 C-terminal segment of the putative C6 protein in V30, site-directed mutagenesis was done to introduce a premature stop codon in place of C6 Lys109. Plants infected with V30ΔC6-d developed similar symptoms to those infected with V30 (Fig. 2.11A; S2 Fig. 2.45). A significant increase in SV compared with V30 was observed at late infection timepoint (32 to 36 dpi) but no change in ULR symptom severity was seen (Fig. 2.11B). Viral load measured at 20 and 36 dpi in V30ΔC6-d-infected plants did not significantly differ from that measured in V30-infected plants (Fig. 2.11C). RT-PCR of total RNA extracted from V30-infected plants was conducted to probe for the presence of a cs strand-derived RNA transcript corresponding to the 3’ region of the putative C6 ORF. C6-specific primers were used with a product being positively identified in RNA extracted from V30-infected plants but not mock-inoculated plants (Fig. 2.11D).

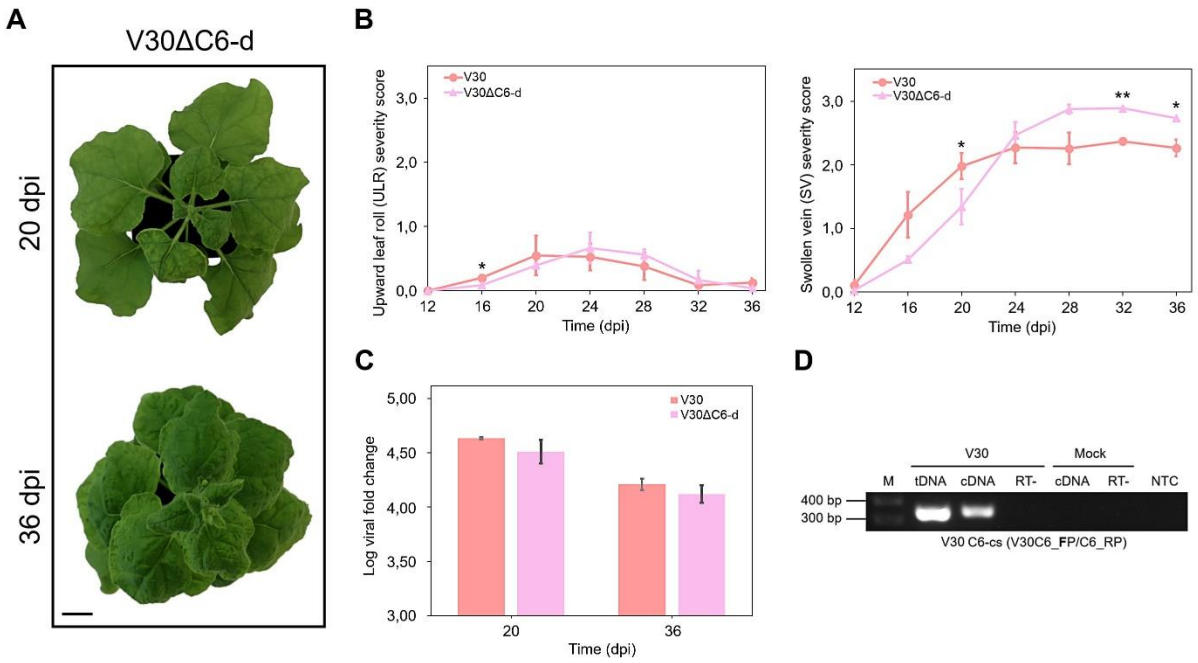


Fig. 2.11 Infection of *Nicotiana benthamiana* with ToCSV V30ΔC6-d infectious clone in single infection. **A**, Symptoms in plants inoculated with V30ΔC6-d at 20 dpi (top) and 36 dpi (bottom). Bar = 2 cm. **B**, Corresponding upward leaf roll (ULR) and swollen vein (SV) symptom severity scores at four-day intervals from 12 to 36 dpi. **C**, Log viral fold change in V30ΔC6-d-infected plants at 20 and 36 dpi. **D**, Agarose gel image of V30 complementary (cs) C6-derived RT-PCR products. M, DNA ladder marker; tDNA, PCR of total DNA extracted from plants inoculated with V30 (positive control); cDNA, RT-PCR of total RNA extracted from plants inoculated with V30 or mock; RT-, negative reverse transcriptase enzyme control; NTC, PCR no template control. Primers used for PCR in brackets with primer used for cDNA synthesis highlighted in bold. Mock is RT-PCR of total RNA extracted from plants inoculated with *Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300. Bars represent mean ± SD. Student's t-test levels of significance: *, P < 0.05; **, P < 0.01.

2.4 Discussion

2.4.1 Investigation in silico of two variant genomes of ToCSV and other African begomoviruses reveals their variable coding potential.

In the present study, the DNA-A coding capacity and in vivo infectivity of single representative isolate DNA-A genome sequences from both variant groups sharing 96 % sequence identity was investigated. The ToCSV-I isolate used was ToCSV-[ZA:Mks30:08] (OK813888) and the ToCSV-II isolate used was ToCSV-[ZA:Mks22:07] (OK813889). Taking the strain demarcation threshold value of 94 % for begomoviruses proposed by Brown et al. (2015) into account, the two ToCSV isolate DNA-A sequences were classified as belonging to one strain group that remains to be

described. Using phylogenetic analysis, it was confirmed that both variant sequences were closely related to solanaceous plant host-infecting begomoviruses from sub-Saharan Africa and the Southwest Indian Ocean islands (Pietersen et al., 2008; Rey et al., 2012; Sande et al., 2021). Pairwise nt sequence comparison of ToCSV-[ZA:Mks30:08] and ToCSV-[ZA:Mks22:07] with isolates belonging to two ToCSV strain groups identified in Mozambique (Sande et al., 2021) indicated that ToCSV-[ZA:Mks30:08] could be placed into the ToCSV-Chok I group. ToCSV-[ZA:Mks22:07] does not belong to either strain group since the DNA-A of this isolate shared <94 % sequence identity with the Mozambican isolates. This suggests that distinct strain groups exist in SA and additional surveys are necessary to explore the genetic diversity of ToCSV in SA.

The two ToCSV variant (V30 and V22) DNA-A genome sequences were subjected to ORF finder analysis. Using BLAST, the six canonical ORFs found in other monopartite begomoviruses were identified namely V1, V2, C1, C2, C3, and C4 (Fiallo-Olivé et al., 2021; Zerbini et al., 2017). The V2 ORF is discussed further in Section 2.4.4. The Rep protein of V30 (encoded by C1) is described in Chapter 3. Traditionally recognised geminivirus ORFs have been assigned an arbitrary protein molecular mass cutoff value of 10 kDa (Padidam et al., 1995; Tan et al., 1995; Torres-Pachero et al., 1993). The smallest currently accepted begomovirus ORF is C4/AC4, which is nested within C1/AC1 (Deom et al., 2021; Medina-Puche et al., 2021). The C4/AC4 ORF is predicted to encode a protein with a molecular mass ~10 kDa and presumably serves as the baseline for the reported cutoff. A serious investigation of putative ORFs present in TYLCV DNA-A encoding proteins smaller than 10 kDa was recently undertaken by Gong et al. (2021). The authors investigated six small putative ORFs encoding proteins with molecular mass ~7 kDa and presented evidence for their existence. To expand coding sequence analysis of ToCSV, ORFs predicted to encode proteins ≥40 aa were included in the present study to examine predicted proteins similar in size to those identified by Gong et al. (2021).

In addition to the canonical ORFs that were identified for the two ToCSV variants, two putative vs ORFs (V3 and V4) and three putative cs ORFs (C5, C6 and C7) were found. The C5 and C6 ORFs are discussed in greater detail in Section 2.4.5. A BLAST search of putative ToCSV V3, V4 and C7 proteins did not yield any significant matches which indicates that these proteins have not been previously annotated. In the current

study, less than 50 % of African begomovirus isolates were predicted to contain V3, V4 or C7 ORFs. A series of small non-canonical ORFs present in the DNA-A of TYLCV were recently described by Gong et al. (2021). None of the small proteins encoded by three novel vs ORFs investigated by the authors were found to align with the ToCSV V3 or V4 proteins. Using TargetP – 2.0 (Armenteros et al., 2019), the V3 protein of both V30 and V22 was predicted to contain SP whilst V4 was predicted to contain mTP. These predicted features could mean that these proteins are translocated within the cell (to the mitochondria in the case of V4) to potentially execute specific functions related to viral infection at different locations within the cell (Duffaud et al., 1985; Jain and Chugh, 2016). Since these rare putative ORFs have not been described previously, their existence and whether they are transcribed remains to be confirmed.

The C7 ORF was found only in African begomoviruses with monopartite genomes in the present study which is unusual since all the other ORFs identified were found in both mono- and bipartite begomoviruses. The ToCSV V30 and V22 C7 proteins were shown to align with the first 45 aa of the protein encoded by “ORF2” (both at 82 % sequence identity) present in TYLCV (Gong et al., 2021). The TYLCV ORF2 protein was predicted to contain TMH (Gong et al., 2021). Both the V30 and V22 C7 proteins were predicted to contain SP but only V30 C7 was predicted to contain TMH. The presence of these two protein features suggest that C7 could be translocated to regions of the cell where it could function as a membrane-associated protein modulating activity across membrane boundaries (Cuthbertson et al., 2005; Duffaud et al., 1985). By expressing the TYLCV ORF2 protein as a GFP-fusion, Gong et al. (2021) showed that this protein localises to the ER, a membrane-rich organelle. The MP of squash leaf curl virus (SLCuV) encoded by BC1 on the DNA-B of this bipartite begomovirus was found to be associated with the ER to facilitate movement (Ward et al., 1997). The CP, V2 and C4 proteins of monopartite begomovirus are known to act as functional analogues of the bipartite begomovirus MPs (BC1 and BV1) (Kumar, 2019; Poornima Priyadarshini et al., 2011; Rojas et al., 2001). Since the C7 ORF was only found in African monopartite begomoviruses in this study, this suggests that the C7 may also serve to be involved in viral movement perhaps as an accessory protein. It could serve a function analogous to that of BC1/BV1 to facilitate viral movement although it was not found in all monopartite begomovirus isolates which suggests that it may not be an essential protein for virus movement. The nt sequence (~500 bp)

upstream of the TYCLV ORF2 start codon was shown to have promoter activity in vivo which implies that this ORF is probably expressed in cells infected with TYLCV (Gong et al., 2021). Additional experiments should be conducted to confirm the expression of this ORF in monopartite begomoviruses such as ToCSV and TYLCV and to determine if it is involved in the inter- and/or intracellular viral movement pathways.

It now recognised that the coding capacity of geminiviruses, particularly the role of smaller ORFs in infection, has been greatly underestimated (Ahmed et al., 2021; Devendran et al., 2022; Gong et al., 2021; Roumagnac et al., 2021). It is hoped that the preliminary investigation of small putative ORFs present in ToCSV and other African begomoviruses described in this study will help to inspire additional work aimed at deciphering their functions in viral replication in the host. Additional sequence analysis of other OW begomoviruses as well as NW begomoviruses will do much to improve the understanding of the occurrence and relevance of these putative ORFs.

2.4.2 Two variant agroinfectious clones of ToCSV induce significantly differing pathologies in *N. benthamiana*.

Agroinfectious clones of ToCSV-[ZA:Mks30:08] and ToCSV-[ZA:Mks22:07] were constructed previously (Esterhuizen, 2012) and are referred to as V30 and V22, respectively. The V30 infectious clone was shown to induce severe leaf curling and plant stunting in tomato cv Rooi Khaki, whilst V22 induced similar but significantly milder symptoms (Esterhuizen, 2012). The present study is the first to report on the infectivity of two infectious clone constructs of ToCSV in *N. benthamiana*. Both V30 and V22 induced virus-like symptoms in this experimental host. The pathologies induced by the two variants differed significantly and were primarily differentiated by the presence of moderate to severe ULR in V22 which was absent in V30. In addition, a recovery in disease symptoms by late infection timepoint was seen in plants infected with V22 but not V30. The sustained SV symptom severity in V30 was correlated with significantly higher viral titre than what was measured in V22-infected plants. Comparison of viral load in tomato cv Rooi Khaki infected with V30 and V22 using Southern blot analysis did not reveal any apparent difference in viral titre although the viral DNA levels were not quantified (Esterhuizen, 2012) as was previously done (phosphorimager analysis) for *N. benthamiana* inoculated with two strains of ToLCNDV (Chatterji et al., 1999). Additional experiments to quantify viral load in

tomato infected with V30 or V22 are required before conclusions can be made regarding the differences in ToCSV variant viral load between the two hosts.

It is apparent that based on viral titre as well as sustained severity of disease symptoms, V30 may be designated the “severe” variant in *N. benthamiana* whilst V22 may be designated as the “mild variant”. If one were to assign variant severity based on the ULR disease symptom, V22 would be described as the “severe” variant whilst V30 would be designated the “mild” variant. Demarcation of geminivirus strain or variant groups on the basis of observed differences in disease phenotype (i.e., “mild” and “severe”) may misrepresent the significance of taxonomical differences between such groups. It has been previously reported that sequence differences of only a few nt may result in significantly altered disease symptom phenotypes and viral accumulation in other geminivirus isolates (Boulton et al., 1991; Brown et al., 2000; Chatterji et al., 1999; Fauquet et al., 2008). These nt sequence differences may induce changes to the aa sequence of proteins encoded by viral ORFs and as little as one aa substitution can significantly alter disease phenotype severity as well as viral tire in infected plants (Chatterji et al., 1999). These differences in observed disease severity are not always modulated by changes to viral protein aa sequence. Analysis of two variants of maize streak virus (MSV) revealed that both symptom severity and host range were dependent upon the sequence identity of two nt which were not implicated in changing viral protein aa sequences (Boulton et al., 1991). These reports indicate that identifying specific sequences responsible for differing symptom phenotypes may prove a challenging task and even more so when attempting to decipher the mechanisms behind these changes (in relation to host-pathogen interactions).

The two ToCSV variant groups (ToCSV-I and ToCSV-II) identified by Esterhuizen (2012) were differentiated by the presence of a recombinant fragment present in the coding region spanning part of the 3’ IR and the first half of the V2 ORF (nt 109-322) in ToCSV-II isolates (including V22). Although V30 and V22 share 96 % whole genome sequence identity, the two variants 79 % sequence identity when comparing the region spanning the recombinant fragment. In addition, V30 contains one additional nt in the 3’ IR region upstream of the canonical V2 start codon which results in the V2 ORF having a putative V2 5’ ORF extension encoding 30 aa resulting in an extended N-terminus (described in Section 2.3.3). Sequence differences present in the IR and V2

coding region of V30 and V22 were identified and their role in the differing pathologies were investigated by generating virus mutant chimeras. The results of these modifications are discussed in Sections 2.4.3 and 2.4.4.

Using RT-PCR, long viral RNA transcripts derived from both vs and cs strands were identified in V30-infected *N. benthamiana*. Two shorter vs transcripts (~500 nt) were mapped to the V2/V1 and V3/V4 coding regions. Two shorter cs transcripts (~500 nt) were mapped to the C5/C6/C7 and C1/C2/C3 coding region overlaps. The presence of transcripts derived from the putative vs and cs ORF coding regions suggests that these ORFs could be transcribed although 5' and 3' rapid amplification of cDNA ends (RACE) assays are necessary to identify the ends of these transcripts. In addition, regions upstream of the start codons of these putative ORFs should be tested for promoter activity which could indicate whether these ORFs are likely to be expressed in vivo during infection.

The longer vs and cs transcripts (~1.2 kb) were mapped to the coding region spanning C1 3' end to V1 5' end. The presence of a long transcript crossing the boundaries of recognised polycistronic mRNA transcripts (Akbar et al., 2012; Mullineaux et al., 1993; Shivaprasad et al., 2005), as well as transcripts being derived from both strands at vs and cs ORF regions may constitute evidence of bidirectional readthrough transcription in ToCSV. This results in the production of long RNAs which span the entire genome including the non-coding IR. Both vs and cs transcripts are produced, and these anneal to each other or to viral mRNAs to produce dsRNA (Pooggin, 2013). Although dsRNA does not function as a replication intermediate in geminiviruses, it is known that these long dsRNA readthrough transcripts are important sources of virus-derived siRNAs (vsRNAs) (Akbergenov et al., 2006; Pooggin, 2013). These vsRNAs are implicated in downstream RNA silencing pathways in other monopartite and bipartite geminiviruses (Akbergenov et al., 2006; Areger et al., 2012; Miozzi et al., 2013; Pooggin, 2013; Rodríguez-Negrete et al., 2009; Yang et al., 2011). Deep sequencing of vsRNAs in TYLCV-infected tomato revealed that vsRNAs were derived from both full genome dsRNA strands and specific regions of viral ORF overlap served as the major sources of vsRNAs (Piedra-Aguilera et al., 2019). Major vsRNA "hotspot" regions in TYLCSV have been mapped to parts of the ORF overlap regions C1/C4 and V2/V1 (Miozzi et al., 2013). The current study provides evidence of long genome-derived transcripts spanning multiple ORFs which could be implicated in RNA

silencing pathways in *N. benthamiana* infected with ToCSV. Further experiments are necessary to confirm vsRNA biogenesis from the identified long RNA transcripts in V30-infected *N. benthamiana*. Deep sequencing of any vsRNAs derived from these transcripts and comparison of vsRNA profiles for the two ToCSV variants in *N. benthamiana* may provide additional clues as to the role that nt sequence differences in the V2/V1 “hotspot” region play in host-pathogen interactions which influence symptom development. Since this region spans the latter part of the recombinant fragment in V22, it would be worthwhile to compare vsRNA profiles between these two ToCSV variants.

2.4.3 The 3’ IR of ToCSV acts as a symptom phenotype determinant in *N. benthamiana*.

The 3’ IR was a prime target for generation of V30 partial sequence swap mutants since the two variant sequences contain several nt differences in this region. In addition, part of the recombinant fragment present in V22 was previously mapped to nt 109 to 137 of the 3’ IR (Esterhuizen, 2021). A significant change in disease symptom phenotype was observed in *N. benthamiana* infected with either V30ΔIR-s or V30ΔIR-sr. This change was more significant in plants infected with V30ΔIR-s than V30ΔIR-sr. The observed development of the ULR is characteristic of V22 infection since this symptom phenotype is not prevalent in wild type V30-induced disease. Although ULR was strongly enhanced in these two V30 IR swap mutants with an associated increase in viral titre at 28 dpi in V30ΔIR-s, symptom recovery typical of V22 infection was not evident in plants infected with these mutants. These observed changes strongly suggest that the sequence identity of the 3’ IR region plays an important role in the symptom phenotype determination in infected *N. benthamiana*. The V22 3’ IR sequence is implicated in the development of ULR irrespective of variant sequence differences present elsewhere in the genome. The absence of disease recovery implies that sequence differences present outside of the IR are involved in mediating recovery phenotype in *N. benthamiana* infected with V22. Although swapping the sequence to the left of the TATA box alone did not alter the symptom phenotype, the significantly increased changes seen in plants infected with the V30ΔIR-s mutants (both left and right sequences swapped) compared to plants infected with V30ΔIR-sr

(sequence to the right of the TATA box swapped) suggests that the identity of the sequence to the left of the TATA box may contribute to the ability of the sequence to the right of the TATA box to act as a symptom determinant.

The hypothesis that the sequence identity of the 3' IR, particularly the sequence to the right of the TATA box, exerts a strong influence on determining symptom phenotype was supported by observations recorded in *N. benthamiana* infected with V22 IR swap mutants Brigden (2022). A near complete knockout in ULR and a reduction of SV symptom severity was observed in plants infected with the mutant containing the ToCSV V30 IR sequences on either side of the conserved TATA box region. No significant change in viral titre was measured in plants infected with this mutant and V22-type symptom recovery phenotype was retained in all mutants (Brigden, 2022). The ULR phenotype was not reduced as significantly in the swap mutant containing the V30 IR fragment to the right of the TATA box than the longer sequence swap mutant (Brigden, 2022). The results obtained by Brigden (2022) also suggest that the presence of the V30 IR fragment does not alter the disease recovery phenotype and that other V22 sequence elements present in regions outside of the IR are responsible for mediating recovery.

The IR is located between the C1 and V2 ORFs in OW begomoviruses and contains the highly conserved stem-loop nonanucleotide sequence where the Rep protein cleaves to initiate rolling circle replication (Pooggin 2013) Several promoter regions that drive bidirectional transcription of downstream viral ORFs, as well as cis-acting elements involved in regulating their transcription are present in this region (Borah et al., 2016; Kumar and Shivaprasad et al., 2020). From multiple sequence alignment of selected African begomovirus IR regions sequences corresponding to key recognised regulatory elements involved in transcription of late vs genes were identified including the CLE motif and the TATA box promoter (Borah et al., 2016; Kumar and Shivaprasad, 2020; Eagle and Hanley-Bowdoin, 1997; Khan et al., 2015). Both ToCSV variant sequences contained an identical single CLE-like motif present in a conserved region of the IR upstream of the region of interest and therefore the sequence identity of this element was not implicated in the changes observed.

No known *cis*-acting elements were identified in the variable region to the left of the TATA box. The variable region to the right of the TATA box in both variant sequences

was shown to contain a TACE. This newly described quasi-palindromic DNA sequence was identified in the 3' IR of isolates belonging to several OW and NW geminivirus genera using multiple sequence analysis and was shown to have the consensus sequence ACTT-(N_7)-AAGT (Cantú-Iris et al., 2019). The authors observed that the N_7 spacer present between the left and right palindromic motifs displayed some sequence variability with a relaxed consensus sequence GGKCCCY. In OW begomoviruses, the TACE is located immediately downstream of the TATA box (Cantú-Iris et al., 2019).

Multiple IR sequence alignment showed that the TACE of the ToCSV-[ZA:Mks30:08] isolate (V30) differed from that of the [ZA:Mks22:07] isolate (V22) and other African begomovirus isolate sequences. Unlike the V22 TACE, the V30 TACE did not conform to the arrangement typical of other OW begomoviruses. The V30 TACE contained a shorter central spacer (N_5) and a GGGAA sequence was situated between the TATA box and the TACE left arm (ACTT). This is the first report of these atypical features in the TACE of any OW begomovirus. It may be worthwhile to generate additional V30 IR mutants which contain either the N_7 spacer sequence of V22 or have a deletion of the GGGAA sequence between the TATA box and the TACE left arm. Such targeted mutation could provide clues as to which of these specific TACE features are responsible for the differing symptom phenotypes and what role these features may exert on transcription regulation activity.

The unusual TACE of V30 makes it a prime candidate for additional experiments to elucidate the role of this novel element in begomovirus-induced disease. Although the changes in symptomology seen in plants infected the V30ΔIR-sr mutant were significant, these changes were more pronounced in the longer IR sequence swap mutant (V30ΔIR-s). It may be possible that an association between the TACE and an unidentified element to the left of the TATA box might exert a combinatory effect to drive specific regulatory activities involved in symptom phenotype development. Therefore, it would be worthwhile to identify specific sequences in the left sequence which contribute to this proposed combinatory effect. Quantitative RT-PCR should be conducted to determine if V1/V2 transcript levels differ between the two wild types as well as selected IR swap mutants. This will reveal whether changes in TACE sequence are associated with altered transcription of vs ORFs.

With the use of strand-specific primers, Yang et al. (2019) previously demonstrated that the IR of TYLCV (nt 2616-147) is transcribed in both directions. In the current study, IR transcripts crossing the Ori (nt 2722-92) being derived from both vs and cs strands were identified in *N. benthamiana* infected with V30 using RT-PCR with strand-specific primers. The presence of these overlapping transcripts strongly suggests that the IR is a potential source of dsRNA which may be processed further to generate vsRNA for RNA silencing pathways (Pooggin, 2013; Piedra-Aguilera et al., 2019; Yang et al., 2019), although the coding regions of geminiviruses tend to yield more significant amounts of siRNAs (Pooggin, 2013). Expression of the TYLCV IR using a tobacco rattle virus (TRV)-based vector system established that the IR acts as a symptom determinant in a susceptible tomato cultivar (Yang et al., 2019). The authors also determined that a short segment of the TYLCV IR to the left of the nonanucleotide (nt 2730 to 2754) showed near total complementarity with a long noncoding RNA (lncRNA) in a susceptible tomato cultivar. Deep sequencing of small RNAs present in infected tomato showed that this 25-nt IR segment was an important source of vsRNA. Evidence for these vsRNA acting as effectors involved in the silencing of susceptible host lncRNA was presented. Interestingly, TRV-based silencing of the target lncRNA designated SlLNR1, resulted in the development of virus-like symptoms (Yang et al., 2019). It is not known why the authors selected SILNR1. BLAST analysis of the 25-nt TYLCV fragment was conducted in the current study and multiple near-perfect matches to several targets present in the tomato genome were found suggesting that multiple host lncRNAs may have been targeted in the study by Yang et al. (2019). Whether the identified ToCSV IR transcripts act as a source of vsRNA which behave as effectors that contribute to pathogenicity is of great interest. Using BLAST, no ToCSV V30 or V22 IR segments with any significant sequence length complementary to that of *N. benthamiana* or the tomato host genome sequences were found. Although no potential host targets were identified, it cannot be ruled out that targets with mismatches may also serve to act as efficient silencers of host RNAs. In any case, infectivity trials of tomato using the various IR swap mutants should be done to determine whether the disease phenotype changes observed in *N. benthamiana* will also be induced in tomato. Preliminary infections of tomato cv Moneymaker with V30ΔIR-s resulted in a visibly milder symptom phenotype (data not shown), although additional experiments are necessary to conclude whether these changes are significant.

Whilst the nt sequence identity of the 3' IR upstream of the canonical V2 start codon has been demonstrated to be vital in mediating symptom phenotype in *N. benthamiana*, the unusual V30 V2 putative 5' coding region which extends into the 3' IR should also be considered. The V30ΔIR-s and V30ΔIR-sr mutants containing the V22 IR fragment to the right of the TATA box lacked the putative V2 N-terminal segment and induced significant ULR symptom phenotypes in *N. benthamiana*. Conversely, V30 and V30ΔIR-sl which contain the corresponding IR region from V30 were predicted to encode the N-terminal V2 extension and induced significantly reduced or no ULR symptoms. Whether the putative N-terminal extensions (or lack thereof) contribute to the observed symptom phenotype changes remains to be investigated. This putative V2 region is discussed further in Section 2.4.4.

2.4.4. The V2 ORF of ToCSV is implicated in modulating disease severity and subsequent symptom recovery in *N. benthamiana.*

The presence of a recombinant fragment in the coding region spanning part of the 3' IR and the first half of the V2 ORF of V22 (Esterhuizen, 2012) makes this a region of interest in examining the influence of sequence differences on symptom phenotype. A preliminary inoculation of tomato cv Rooi Khaki with a V30 V2 ORF swap mutant (canonical V30 V2 ORF replaced with V22 ORF sequence) did not result in a change in symptom severity (Esterhuizen, 2012). Southern blot analysis demonstrated that apparent levels of viral DNA in infected plants did not differ compared to plants infected with the wild type variants although viral titre was not quantified (Esterhuizen, 2012). In the present study, the V30 V2 ORF swap mutant (V30ΔV2-s) and a V22 canonical V2 ORF swap mutant (V22ΔV2-s) were used to examine the possible role of sequence differences in this region in the differing pathogenicity of V30 and V22 in *N. benthamiana*.

The V2 ORF swap mutants did not display altered disease symptom phenotypes in *N. benthamiana* as was observed with the IR swap mutants. Instead, the key difference recorded was a significant change in disease severity and altered levels of symptom recovery. Plants infected with V30ΔV2-s mutant showed a significant recovery in SV symptoms by late infection timepoint accompanied by significant reduction in viral load in comparison with V30-infected plants which do not display symptom recovery. The

V22ΔV2-s mutant induced severe ULR and SV symptoms which persisted until late infection timepoint with significantly increased viral load at peak and late infection, signifying an effective knockout of the wild type V22 recovery phenotype. These results confirmed the previous assertion that a region outside of the IR was responsible for the presence of the V22 recovery phenotype.

When identifying the precise sequence elements responsible for these observed changes in disease severity and recovery it is important to consider all the coding and non-coding elements present in the V2 ORF region that was swapped. This region includes a V1/V2 ORF overlap which is a known “hotspot” that acts as a major source of vsRNA in both TYLCV and TYLCSV (Miozzi et al., 2013; Piedra-Aguilera et al., 2019). Recovery in geminivirus-infected plants has been linked to the presence of vsRNAs involved in RNA silencing pathways (Chellappan et al., 2004; Rodríguez-Negrete et al., 2009). Differences in pathogenicity or infectivity observed in this study between variant isolates may thus be mediated by the differing vsRNA profiles which result from the biogenesis of these vsRNAs from genomic regions containing nt sequence differences. In addition, variant-specific putative promoter, cis-acting elements and possible iteron sequences present in this region could also be responsible for modulating expression of viral ORFs. Although promoter elements have not yet been mapped for ToCSV, sequences with promoter activity have been found within the coding regions of other begomoviruses (Gong et al., 2021; Wang et al., 2013, 2014a). Promoter mapping of the two variant sequences may uncover additional sequence features which are implicated in differentiating pathogenicity.

By swapping the canonical V2 ORF, overlapping portions of other ORFs present in this sequence were also swapped. Consequently, changes in the aa sequence of the proteins encoded by these overlapping ORF fragments occurred. Since the putative ToCSV C5, C6 and C7 proteins have not yet been fully characterised, the relevance of aa sequence differences in selected regions of these proteins cannot currently be established. The sequence identities of two aa in the N-terminal region of the CP (aa 36 and aa 55) were swapped in the V2 ORF swap mutants. Changes to this region could alter the activity of the CP. The N-terminal region (aa 1-62) of tomato yellow leaf curl Thailand virus (TYLCTHV: AF141922) was shown to be essential for the ability of CP to bind both dsDNA and ssDNA (Pitaksutheepong et al., 2007). The CP of ACMV (J02057) was experimentally shown to contain a NLS at aa 1-54 (Unseld et al., 2001).

Using CP deletion mutants, an RKPR sequence motif present at CP aa 52-55 of ToLCJaV (AB100304) was shown to be critical for the ability of the CP to localise to the nucleus (Sharma and Ikegami, 2009). The sequence of this NLS in the CP of V30 and V22 is RKPK and RKPR respectively. The substitution at CP aa 55 in V2 swap mutants could potentially alter the activity of the NLS leading to modified virus-host interactions which influence the development of disease symptoms. Targeted substitution of CP aa 55 via generation of substitution mutants may prove useful in determining whether the sequence difference at this position contributes to the observed differences in symptom phenotype in *N. benthamiana*.

The critical roles that the V2/AV2 protein of both mono- and bipartite geminiviruses plays in viral infection and pathogenicity has been reviewed extensively (Csorba et al., 2015; Kumar, 2019; Kumar and Shivaprasad, 2020). The V2/AV2 protein is a known suppressor of PTGS (Basu et al., 2018; Sharma and Ikegami, 2010; Zhang et al., 2012; Zrachya et al., 2007). Disease recovery which typically manifests as symptom remission in geminivirus-infected plants has been linked to various RNA silencing pathways (Chellappan et al., 2004; Rodríguez-Negrete et al., 2009). The change in disease recovery levels seen in plants infected with the V2 ORF swap mutants therefore warranted greater scrutiny of the sequence differences present in the two variant V2 proteins. A portion of the V22 recombinant fragment (nt 109-322) encodes the first 61 aa of the V2 protein. The N-terminal region of both ToCSV variant isolate V2 proteins contained seven aa differences. The predicted ss of this region included three helices. No solved structure for any geminivirus V2/AV2 protein currently exists but the predicted ss of ToCSV V2 is comparable to that predicted for other begomovirus V2 proteins (Li et al., 2021a; Poornima Priyadarshini et al., 2011). Four of the V2 aa sequence differences were located to the first helix and the unconserved residue at aa position 47 was also shown to be highly variable amongst African monopartite begomovirus V2 proteins. In another study of ToCSV, inoculation of *N. benthamiana* with V22 V2 Val27Sel and Thr58Ser substitution mutants (V22 residue mutated to V30 residue) resulted in significant changes to the V22-induced disease severity and progression (Brigden, 2022).

Plants infected with the V2 Val27Ser mutant developed significantly increased ULR and SV symptoms compared to V22-infected plants (Brigden, 2022). This change was comparable to that induced by V22ΔV2-s in the present study although a slight

recovery in symptoms was observed in plants infected with the Val27Ser mutant (Brigden, 2022). Viral titre was significantly increased at late infection but not peak infection timepoint. A significant delay in ULR and SV symptom development was observed in plants infected with the V2 Thr58Ser mutant but no change in viral load or peak symptom severity was reported. Disease recovery was not affected by this mutation (Brigden, 2022). The results obtained by Brigden (2022) suggest that sequence differences present in the V2 N-terminal region, particularly aa position 27, exert a significant influence in modulating ToCSV-induced disease severity as well as symptom recovery in *N. benthamiana* infected with V22. Comparison with V22ΔV2-s suggests that although V2 aa sequence identity is important, additional sequence differences are implicated in the more significantly increased disease severity and viral titre observed in the full V22 V2 ORF swap mutant.

Although no published information exists regarding the specific function of V2 aa 27, this residue is flanked on either side by hydrophobic regions containing residues which have been shown to be necessary for pathogenicity and efficient RNA silencing activity in other geminiviruses (Li et al., 2021a; Luna et al., 2020). Two tyrosine residues at aa positions 24 and 32 were required for efficient silencing suppression activity of TbCSV V2 protein (Li et al., 2021a). These two tyrosine residues were conserved in all African begomovirus isolate V2 proteins investigated in this study. Their proximity to the short variable motif (which includes aa 27) might moderate their specific function through changes in molecular structure or positioning of other key residues involved in protein binding. Since no solved structure of any geminivirus V2/AV2 protein exists, the arrangement of any putative binding pockets or catalytic regions must be elucidated via structural studies before any details of mechanistic processes involved host-protein interaction involved in RNA silencing suppression activity can be described.

It has been shown that the AV2 of ToLCNDV interferes with the ability of RDR1 to function in the RNA silencing pathway in *N. tabacum* (Basu et al., 2018). In *N. tabacum* infected with ToLCGuV symptom recovery was observed with a corresponding increase in host RDR1 expression levels as well as higher levels of viral promoter region DNA methylation (Basu et al., 2018). The AV2 protein of ToLCNDV was found to function as a stronger PTGS suppressor than ToLCGuV AV2. Despite the two AV2 proteins sharing 74 % aa sequence identity and having identical sequence motifs implicated in silencing suppression, it was suggested that the regions containing

sequence differences may account for the differing activities of the proteins (Basu et al., 2021). This indicates that even if key functional motifs present in the V2 N-terminal region are conserved, the sequence identity of adjacent residues should be considered in comparative V2 protein studies.

One of the sequence differences present in the V30 and V22 V2 N-terminal region included the residue at position 58 which was mapped to the putative PKC motif (aa 58-60). This motif has previously been reported to have a consensus sequence of TXR (Chowda-Reddy et al., 2008; Mubin et al., 2010; Saeed et al., 2018). Alanine substitution of the threonine and arginine residues of the PKC motif previously showed that these residues are required for the subcellular localisation, pathogenicity-enhancing, and RNA silencing suppression activities of several geminivirus V2 proteins (Chowda-Reddy et al., 2008; Li et al., 2021a; Luna et al., 2020). Although this motif was mostly conserved amongst African begomovirus V2 proteins in the current study, a serine in place of the threonine was identified in the PKC motif of V30, ToCSV exemplar isolate (AF261885), and tomato leaf curl Ghana virus (ToLCGV: EU350585) V2 proteins. The presence of a serine at this position has been reported in selected cotton-infecting begomovirus isolates (Luna et al., 2020; Mubin et al., 2010; Saeed et al., 2018) but the effect of substituting the PKC motif threonine with an evidently rarer serine residue has not been previously investigated in other geminivirus V2 proteins. The findings reported by Brigden (2022) suggest that although the V22 V2 Thr58Ser mutation does not impair viral replication in V22, the delay in symptom development could be due to possible interference with V2 function as a result of this atypical PKC motif but further experiments are necessary to confirm this. It is not known what, if any, positive influence a serine at V2 position 58 exerts on pathogenicity in V30 infection. In the present study, no delay in symptom development was seen in V30-infected plants (when compared with V22), suggesting that the non-consensus serine present in V30 PKC motif is not detrimental to V2 function in V30 infection. It is possible that this altered PKC motif may result in modified V2 activity that is apparently advantageous to the pathogenicity of V30 (but not V22) and mutation of this residue to threonine in V30 should be conducted to gain a better understanding of the relevance of this atypical PKC motif residue.

It has been demonstrated that the aa sequence identity of the V2 protein could play a crucial role in the differing ToCSV variant pathologies, and additional experiments are

needed to identify the precise functions of selected aa present in the V2 proteins of V30 and V22 which result in their distinctive pathologies in *N. benthamiana*. Mutation of the V30 V2 ORF is necessary to determine if a V2 Ser27Val substitution will consequently reduce pathogenicity and result in a recovery phenotype as was recorded in plants infected with the V30 full V2-swap mutant. The role of V2 aa 27 in RNA silencing suppression pathways should be investigated further. Substitution mutation to investigate the other sequence differences present in the N-terminal region spanning the three predicted helices is limited by the presence of the overlapping C6 ORF and differences at aa positions 47 and 52 cannot be substituted with the residue of the other variant without altering the predicted aa sequence of the putative C6 protein. Expression of mutated V2 proteins using a PVX vector expression system could be done to compare the effect of mutation on the ability of V2 to act as a pathogenicity determinant. Additionally, two-hybrid screening assays using V2 mutants to investigate specific protein-protein interactions described previously for other begomoviruses are also suggested.

Only the canonical V2 ORF region was swapped in the current study. The role of the putative V30 V2 5' extension in pathogenicity was not investigated. The putative V30 V2 start codon lies upstream of the TATA box/TACE promoter region which suggests that this could be a spurious artefact. The canonical V2 start codon was identified in all African begomovirus V2 ORFs examined in this study and extensions of the V2 5' end beyond the canonical start codon were rarely found in other African begomoviruses. The rarity of such ORF 5' extensions beyond the canonical start codon does not necessarily mean that they do not exist as valid extensions of canonical ORFs. The atypical C4 ORF of an isolate of cotton leaf curl Burewala virus (CLCuBuV) contains a putative start codon upstream of both the C1 start codon and the TATA box promoter located in the 5' end of the IR. The presence of the start codon upstream of the promoter known to drive the expression of cs ORFs (Eagle and Hanley-Bowdoin, 1997) was unexpected. Transcript mapping of *N. benthamiana* infected with CLCuBuV showed that this 5' extended C4 ORF was transcribed as evidenced by the presence of a long transcript (~1.7 kb) containing C4 (Akbar et al., 2012). The authors did not determine whether the extended C4 ORF was translated but did not rule out the existence of an N-terminally extended C4 protein. An isolate of PaLCuV was reported to contain a V2 5' extension encoding a 32 aa N-terminal

segment upstream of the canonical V2 methionine (Mubin et al., 2010). The authors determined that deletion of these putative residues did not alter the ability of V2 to induce a HR in *N. benthamiana* when using a using PVX vector expression system. They concluded that the true start of the PaLCuV V2 protein was likely the “bona fide” methionine (aa position 33) present at the N-terminus of other begomovirus V2 proteins lacking N-terminal extensions (Mubin et al., 2010). The authors did not probe for an extended V2 transcript and did not experimentally determine the true length of the V2 transcript. The lack of substantial published information regarding unusual N-terminal extensions of canonical ORFs renders them a mysterious phenomenon that should be explored further despite their rarity. It is possible that such features could allow for expression of more than one form of a particular protein which could have differing activities. In this study the V2 N-terminal extension was predicted to contain a TMH and SP which could hint at membrane-associated function for this form of the V2 protein.

In the current study, a vs transcript spanning the putative 5’ V2 region, extending into the canonical V2 ORF sequence was detected in V30-infected plants using sequence-specific primers. A transcript spanning the same region in V22 was also identified which was unexpected since V22 is not predicted to encode a V2 N-terminal extension when following standard code. It may be possible that the positive detection of this transcript in V22-infected plants was due to the presence of read-through transcripts spanning the full genome (discussed in Section 2.4.2). Additional experimental approaches combining various techniques such as V30 promoter/transcript mapping, PVX vector-mediated expression and comparative subcellular localisation studies of the two possible forms of the V30 V2 protein are required. Determining the true length of the V30 V2 ORF could shed some light on the function that any putative N-terminal extensions could have in mediating pathogenicity. Such unique sequence features may also require revision of the boundaries of the IR which may not be as defined as previously believed.

2.4.5 C5/AC5 and C6/AC6 – additional players in the game that require further consideration.

Short putative geminivirus ORFs have recently gained increasing attention for their role in viral infection (Ahmed et al., 2021; Gong et al., 2021; Li et al., 2021b; Teixeira et al., 2021). In the current study, ORF finder was used to investigate the genomes two ToCSV variants and 53 African begomovirus isolates. The C5/C5 and C6/AC6 ORFs were the most frequently identified putative ORFs. The protein sequences encoded by these two ORFs were highly divergent and variable in length. Both ORFs are located within the DNA-A region spanning the 3' end of C3/AC3 to the 5' end of the V2/AV2 ORF. The relative arrangement of these two ORFs was consistent amongst all begomovirus isolates analysed. Both were located on the cs strand, C5/C5 was in frame with V1/AV1 and C6/AC6 was in frame with V2/AV2.

With the rapid increase in identified begomovirus isolates and better understanding of their pathology due to the development of more sensitive molecular/data processing techniques, the number of additionally recognised ORFs will inevitably increase. As a result, confusion has arisen regarding their nomenclature. Various conflicts for identified "C5/AC5" and "C6/AC6" ORFs are evident in existing sequence databases. Attempts to align these putative sequences often results in erroneous results due to the inevitable confusion between the highly variable sequences of these two rare ORFs. This is particularly relevant to the C5/AC5 ORF which was found in this study to have a segmented translational profile in most African begomovirus isolates examined. This was confirmed by the presence of long continuous C5/AC5 proteins identified in some isolates which aligned with fragmented translation products (as a result of multiple in frame stop and methionine codons) predicted for other isolates. An alternative hypothesis is that stop codon readthrough might allow for the translation of a long C5/AC5 protein like that encoded by the C5/AC5 ORF of other isolates that do not contain internal in frame stop codons. Stop codon readthrough has been reported in RNA viruses of plants and is known to be mediated by the presence of repeat sequence motifs and ss elements present in the vicinity of the stop codon (Firth and Brierley, 2012; Miras, 2017) This non-canonical translation strategy has not been reported in geminiviruses and was not experimentally investigated in the present study.

In a study by Bigarré et al. (2001), two putative ORFs, termed “AC5” and “AC6”, were annotated for an Egyptian isolate of cotton leaf curl Gezira virus (CLCuGeV). ORF finder analysis of the DNA-A of CLCuGeV conducted in the present study identified these two ORFs as being in frame with each other and AV1. This suggests that the “AC6” designation reported by Bigarré et al. (2001) may be inaccurate and that two discrete AC5 proteins are encoded by the CLCuGeV AC5 ORF. In a study of Indian begomovirus isolates by Ansar et al. (2019), the ORF having codons in frame with AV1 was termed “AC6” whilst that having codons in frame with AV2 was termed “AC5”. ORF finder analysis of these Indian begomoviruses and multiple aa sequence alignment of putative AC5 and AC6 proteins was conducted in this study (data not shown) showed that Indian begomovirus “AC5” proteins aligned with African begomovirus C6/AC6 proteins and Indian “AC6” proteins aligned with African begomovirus C5/AC5 proteins. Such inconsistencies are expected when identifying putative ORFs since their fragmented and highly variable protein product sequences contrast with the accepted ORFs which are highly conserved and show less variability in length. The presence of another putative ORF in the C5/AC5 and C6/AC6 coding region (herein termed C7) will inevitably further complicate matters, although C7 is even rarer than C5/AC5 and C6/AC6. As the number of new ORFs recognised to serve key functions in geminivirus infection increase, a consistent nomenclature system will need to be developed. This may be assisted in part by using multiple sequence alignments and phylogenetic analysis to determine the consensus size and location of novel functional ORFs.

The first major study of several OW and NW geminivirus C5/AC5 proteins by Li et al. (2015), presented evidence for the existence of two recognised C5/AC5 protein domains: “Gemini AC5-1” (pfam04807) and “Gemini AC5-2” (pfam08464). These two domains are herein referred to as AC5-1 and AC5-2. The majority of African begomovirus C5/AC5 proteins contained at least one of these domains. ToCSV C5 proteins contained the AC5-2 domain only. A putative domain, termed the NLHr domain, was present in some African begomovirus C5/AC5 proteins identified in this study and was predicted to contain a long TMH. The NLHr was not as prevalent as the other two domains in African begomovirus C5/AC5 proteins but contained some conserved residues. Its unique structural features warrant additional studies to determine whether it influences the subcellular localisation and activity of the C5/AC5

protein in infected cells. Specific functions of the recognised AC5-1 and AC5-2 domains are not yet fully understood. The precise activity of single conserved residues and sequence motifs present in these domains have not yet been experimentally determined. The C-terminal region of AC5 protein of MYMIV containing the AC5-1 domain was shown to be essential for the ability of the AC5 to suppress TGS in vivo (Li et al., 2015). The aa sequence of the protein encoded by TYLCV “ORF1” (Gong et al., 2021) aligned with African begomovirus C5/AC5 proteins and contained a recognised AC5-1 domain sequence. Sub-cellular localisation studies of the ORF1 protein (herein regarded as C5) fused to GFP revealed that this C5 protein localised to the nucleus. These previously reported results suggest that the TGS suppression activity mediated by residues present in the AC5-1 domain may at least initiate in the nuclei of infected cells. The role of the AC5-2 domain present in ToCSV C5 requires further study since its function is currently unknown. Site-directed mutagenesis of selected residues or deletion of the C5/AC5 ORF in infectious clone constructs of geminiviruses will prove particularly challenging for two reasons. Firstly, considerable overlap exists between multiple ORFs in the C5 coding region increasing the risk of introducing inadvertent mutations to these other ORFs. Secondly, the C5/AC5 ORF contains several internal methionine codons rendering deletion of the entire ORF difficult. Interestingly, these internal methionine codons imply that a repertoire of C5/AC5 proteins could be expressed by begomoviruses. Keeping these challenges in mind in the current study, a C5 deletion mutant of ALCScV generated by Li et al. (2021b) was closely examined. It was discovered that the C5 ORF was only partially deleted in their experiment with 71 C-terminal residues predicted to have remained due to the presence of a start codon downstream of the mutated start site. In addition, an inadvertent mutation was introduced into the C3 ORF resulting in a Trp114Arg substitution in the C3 protein. Since C5 was only partially deleted and an inadvertent mutation was introduced into a canonical ORF, the results obtained by Li et al., (2021b) may have been due to these inadvertent alterations of the ALCScV genome. This example highlights that care that must be taken when investigating the C5/AC5 ORF via generating mutant virus clones.

Whilst AC5/C5 has been shown to function as a pathogenicity determinant and a potent inhibitor of antiviral silencing pathways in *N. benthamiana* (Li et al., 2015, 2021b), the role of AC6/C6 in viral infection remains to be elucidated. This is not

surprising since it was observed in the present study that AC6/C6 occurs more rarely than AC5/C5 in African begomoviruses. The rarity of C6/AC6 has been noted previously in begomoviruses from Southeast Asia (Du et al., 2014). In addition, AC6/C6 homologues in the genomes of begomoviruses from the Americas could not be identified. This implies that like V2/AV2, the AC6/C6 ORF is only present in OW begomoviruses. The sequences of African begomovirus AC6/C6 proteins were divergent and variable in length, although some regions of high sequence conservation were present. The putative C6cd domain was found to contain several conserved residues in African begomovirus C6/AC6 proteins and was identified as a target for future studies to investigate this conserved domain for key residues which might be involved in C6/AC6 activity. The V30 C6 protein (143 aa) was the longest African begomovirus C6 protein identified in this study. The C6 protein of V22 was shorter (108 aa), lacking the 35 C-terminal residues present in V30 C6. Comparison of V30 and V22 C6 protein sequences with proteins encoded by small non-canonical TYLCV ORFs described by Gong et al. (2021) showed that the 62 C-terminal residues of V22 C6 shared 79 % sequence identity with a protein encoded by TYLCV “ORF3”. The TYLCV ORF3 protein (herein regarded as C6) was shown by the authors to localise to the mitochondria in *N. benthamiana* when expressed as a GFP-fusion. Interestingly, both V30 and V22 C6 proteins were predicted to contain SP and a transmembrane domain in the N-terminal region using the Phobius and TargetP – 2.0 webservers (Armenteros et al., 2019; Käll et al., 2007). Gong et al. (2021) did not identify a transmembrane domain in their TYLCV C6 protein and theirs also lacked the N-terminal region containing the transmembrane domain present in V30 and V22 C6 proteins. This suggests the exciting possibility that the C6/AC6 protein may have a repertoire of accessory functions which might serve to fine-tune crucial infection events occurring at different regions within the cell. Such events driven primarily by the more conserved canonical proteins may be assisted in part by the activity of non-canonical proteins such as C6/AC6. Further testing is necessary to determine if the observed variability in C6/AC6 protein length enables diversity in protein function. In addition, the C6 proteins of ToCSV and other African begomoviruses contained two to three internal methionine residues further supporting the hypothesis that multiple forms of this protein may exist.

In this study, the role that the 35 C-terminal residues of V30 C6 might have in viral infectivity in *N. benthamiana* was investigated by generating a C6 C-terminal deletion mutant (V30ΔC6-d). Although viral load was not affected, the severity of the SV symptom phenotype was significantly increased at late infection. From this, it was concluded that the C-terminal region of V30 C6 is not required for successful infection in *N. benthamiana* but may be implicated in modulating symptom severity via an unknown mechanism. Additional mutagenesis experiments designed to mutate several additional C6 regions of interest such as the putative C6cd were severely restricted by the overlapping of V1, V2 and C5 ORFs. Modification of the V22 C6 stop codon to a lysine codon (present at this position in V30 C6) was not possible without inadvertently modifying the N-terminus of the V1 protein. PVX vector expression of wild type and mutant C6 proteins should be conducted to determine if C6 is a pathogenicity determinant and to identify key sequence motifs that could serve critical function in pathogenicity. RT-PCR assays confirmed that the 3' region of the V30 C6 ORF is transcribed and a 494 nt cs transcript containing the entire C6 ORF was also identified. Although these results indicate that the cs strand region spanning the C6 ORF is transcribed in V30-infected *N. benthamiana*, 5' and 3' RACE assays are necessary to identify the ends of the C6 transcript thereby determining its true length. Quantitative RT-PCR to compare C6 transcript expression levels at different timepoints and mapping of putative C6 promoter regions will provide more insight into the expression of C6.

The current study was the first to identify and map the positions of the putative C5, C6, C6, V3 and V4 ORFs in ToCSV V30 and V22. An exhaustive search of these putative ORFs in related African monopartite and bipartite begomoviruses determined that that C5 and C6 are the most prevalent and that C7 is present in monopartite begomoviruses only. The differing pathologies induced by ToCSV V30 and V22 in *N. benthamiana* were characterised for the first time. The two symptom phenotypes were primarily differentiated by the ULR symptom phenotype and the presence of symptom recovery in V22-infected plants. Results of qPCR analysis showed that both variants induced a peak viral load at 20 dpi, with V30-infected plants having significantly higher viral titre than V22-infected plants at peak and late infection timepoints. Using Southern blot analysis (without quantification) no difference in viral titre between the two variants was previously reported in tomato cv. Rooi Khaki

(Esterhuizen, 2012). Additional infectivity trials using a range of tomato cultivars with viral load quantification will allow for more precise comparison of viral titre in the two hosts. The present study presented evidence of long RNAs transcribed from both strands which could potentially be processed to generate siRNAs (Pooggin, 2013) and deep sequencing of small RNAs may assist in confirming this hypothesis. Infection of *N. benthamiana* with V30 IR swap mutant chimeras showed that nt sequence differences present in the 3' IR sequence between the TATA box and the V2 start codon are responsible for the symptom phenotype determination. The two variant IR sequences contain distinct TACE (Cantú-Iris et al., 2019) in this region which are proposed to contribute to the altered disease phenotypes via a mechanism which requires additional experiments to determine. Measurement of vs transcript levels could shed light on differential ORF expression which could result from the influence of the unique TACE features present in V30 3' IR. The observed changes to disease severity and viral titre in plants infected with V2-swap mutants suggest that viral titre as well as the V22 recovery phenotype is modulated by V22-specific sequence present in this region. Closer examination of the aa sequences of the putative African begomovirus C5/AC5 proteins led to the discovery of a third putative domain termed the NLHr which was predicted to contain a transmembrane domain. A conserved domain designated C6cd was identified in C6/AC6 proteins. The function of these domains and important residues involved in the activity of these two proteins requires further attention since these putative ORFs are gaining greater recognition for their role in geminivirus infection (Gong et al., 2021; Li et al., 2021b). Through generation of a C6 deletion mutant it was demonstrated that the 35 C-terminal aa of V30 C6 (absent in V22 C6) are not required for infectivity. The presence of RNA transcripts derived from the C6 coding region indicates that this ORF is transcribed. The current work is aimed at improving the understanding of the complex dynamics involved in ToCSV pathogenicity. Many possible avenues of future investigation have been identified in this ground-breaking study of ToCSV pathogenicity in *N. benthamiana*. It is hoped that these findings will ultimately assist in the aim of developing suitable strategies intended for the surveillance and control of this pathogen in SA.

Chapter 3: ToCSV Rep is predicted to function as an HUH endonuclease and an ATPase/superfamily 3 helicase

3.1 Introduction

Viruses belonging to the ssDNA virus family *Geminiviridae* pose a serious threat to the global cultivation of a wide variety of important crop and horticultural plant species from the dicotyledonous and monocotyledonous angiosperm groups. This is in part fuelled by the rapid proliferation of their insect vectors (Rojas et al., 2018). With the recent improvements in DNA sequencing technology and methods associated with environmental sample genomics, the number of recognised genera has increased to 14 which includes over 500 recognised species, the majority belonging to the genus Begomovirus (Fiallo-Olivé et al., 2021; Roumagnac et al., 2021). The stem-loop element containing the Ori is common to all geminiviruses and includes the nonanucleotide sequence (5'-TAATATTAC-3') which is conserved in nearly all recognised species except those belonging to the genera *Eragrovirus* and Becurtovirus which contain the nonanucleotide sequence 5'-TAAGATTCC-3' (Fiallo-Olivé et al., 2021).

This region of the vs strand serves as the DNA cleavage site for Rep to initiate RCA (Akbar et al., 2012; Piedra-Aguilera et al., 2019). Rep is encoded by all geminiviruses and is responsible for recruiting and coordinating host polymerases and associated factors to drive viral replication in the host cell nucleus (Rizvi et al., 2015; Ruhel and Chakraborty, 2019; Hanley-Bowdoin et al., 2013). This protein has also been demonstrated to execute several important enzymatic activities at various stages of replication. Upon binding to the viral genome dsDNA intermediate, Rep nicks the viral-sense strand at the Ori (endonuclease activity). By catalysing the hydrolysis of ATP (ATPase activity), Rep is then able to unwind the dsDNA (helicase activity). Following vs strand extension mediated by host DNA polymerases, Rep cleaves the newly synthesised vs strand and ligates the free ends (ligase activity) to generate a circular ssDNA (Clérot and Bernardi, 2006; Hanley-Bowdoin et al., 2000; Pooggin, 2013; Yoon-Robarts et al., 2004). Important residues implicated in the enzymatic activity of geminivirus Rep have been mapped to two distinct protein domains containing functional motifs conserved amongst diverse DNA virus Rep proteins with similar

functions. These two domains include the N-terminal HUH endonuclease domain and a C-terminal ATPase/SF3 helicase domain (Desbiez et al., 1995; George et al., 2014; Orozco et al., 1997; Tarasova and Khayat, 2022). A short oligomerisation domain separating the two enzymatic domains is necessary for the oligomeric assembly of Rep monomers (Choudhury et al., 2006; Orozco et al., 1997). A fourth domain, the C-terminal domain, has been identified in TYLCSV Rep but its precise function has not been determined (Clérot and Bernardi, 2006).

Whilst the molecular structures of other viral Rep proteins in monomeric and oligomeric form have been solved previously, limited structural information is available for geminivirus Rep proteins. Only the monomeric structures of the N-terminal endonuclease domains of TYLCV and WDV Rep proteins have been solved and described (Campos-Olivas et al., 2002; Everett et al., 2019). Three conserved functional motifs (motifs I, II, III) involved in DNA binding, cleavage and ligation have been mapped to catalytic region containing an α-helix and an antiparallel β-sheet comprised of five β-strands. This arrangement has also been identified in the structures of other ssDNA virus Rep N-terminal domains (Campos-Olivas et al., 2002; Everett et al., 2019; Vega-Rocha et al., 2007a, 2007b). The iteron-related domain (IRD) and the GRS are two additional sequence motifs that have been identified in geminivirus Reps and the role of these motifs in the endonuclease domain function has been experimentally determined (Argüello-Astorga and Ruiz-Medrano, 2001; Chatterji et al., 1999; Nash et al., 2011).

Although no solved structure of the geminivirus Rep C-terminal ATPase/SF3 helicase domain exists, this domain contains the conserved sequence motifs found in other SF3 helicases. Homology modelling of this domain has been used to demonstrate the function of specific motif residues identified to be crucial to enzymatic activity (George et al., 2014; Ruhel et al., 2021). These motifs include WA, WB, B’ motif, and C motif which contain conserved residues that have been demonstrated to serve important functions in ATP hydrolysis coupled with DNA unwinding activity (Desbiez et al., 1995; Hickman and Dyda, 2005; Koonin, 1993). Although the active form of other viral SF3 helicases is known to exist as hexamers or heptamers (Hickman and Dyda, 2005; Santosh et al., 2020; Tarasova et al., 2021), the helicase activity of Rep is reliant upon the assembly of Rep monomers into higher order oligomers ranging from dodecameric to 24-meric structures (Clérot and Bernardi, 2006; Choudhury et al., 2006). This

represents a unique functional conformation not reported for other SF3 helicases. The helicase domain of geminivirus Rep is unique in that it lacks the arginine finger motif which is responsible for coordinating ATPase activity linked to conformational changes and mediating interactions between adjacent monomeric units in other viral SF3 helicases (Gai et al., 2004; Moreau et al., 2007). Attempts to solve the structure of geminivirus Rep helicase domain is the subject of ongoing study since its unique characteristics may serve to redefine current models of how SF3 helicases function at the molecular level.

In silico investigation of the aa sequence of the Rep proteins of ToCSV and related African begomovirus Reps in the present study revealed that these Rep proteins contain conserved functional motifs involved in endonuclease and helicase activity. Homology modelling of the two important catalytic domains of V30 Rep revealed a structure similar to that of other viral Rep proteins in terms of ss arrangement and relative positions of functional motifs. Using C1 truncation mutants the 14 C-terminal residues of Rep were shown to not be essential for infectivity in *Nicotiana benthamiana* although disease symptom development was significantly delayed, and viral titre was reduced. A recent study suggested that the C-terminus of BPV Rep containing a disordered acidic region could be implicated in stabilising the helicase structure (Javed et al., 2021). Recombinant expression of V30 Rep helicase and C-terminal domains (RepC) in *Escherichia coli* proved challenging and several expression and purification experiments indicated that further optimisation was necessary to improve protein purity and solubility. C-terminal truncation of RepC is suggested as a rational approach to achieve this. This study has provided a significant base for describing key V30 Rep residues which are necessary for Rep activity and could be targeted for mutation to investigate their role further. In addition, regions of the protein which could be targeted for removal to ultimately improve chances of obtaining pure protein in amounts significant enough for future structural studies have also been identified.

3.2. Methods and Materials

3.2.1 ToCSV wild type and C1 mutant infectious clone constructs

The agroinfectious clone construct of the wild type severe ToCSV variant ToCSV-[ZA:Mks30:08] DNA-A (OK813888) used in this study is termed “V30” and is described

in Chapter 2, Section 2.2.1. Two point mutation infectious clone constructs were generated via site directed mutagenesis using V30 as the template. The 3' end of the C1 ORF encoding the C-terminal domain of the Rep protein was selected as a target for mutagenesis to truncate the C-terminal end of this domain. Mutagenesis targets were carefully selected to avoid the introduction of inadvertent mutations to the overlapping canonical C2 ORF. The two point mutants included a 14 aa C-terminal deletion mutant [V30ΔC1-d(14)] and a 22 aa C-terminal deletion mutant [V30ΔC1-d(22)].

To generate V30ΔC1-d(14), a premature stop codon was introduced in place of the codon for canonical Rep Gln346. Mutagenesis was conducted using the QuikChange® II XL Site-Directed Mutagenesis Kit (Agilent Technologies). A fully overlapping mutagenesis primer pair was designed using the QuikChange® Primer Design Program web server (https://www.agilent.com/store/primerDesignProgram.jsp) in accordance with recommended guidelines. Mutagenesis primers were synthesised and purified using polyacrylamide gel electrophoresis (PAGE) (Inqaba Biotech). Mutagenesis PCR with *PfuUltra* High-Fidelity DNA polymerase was set up as per manufacturer's instructions using 10 ng V30 template DNA. A three-step cycling protocol was followed with 18 cycles and an extension time of 720 s (60 s/kb). Following PCR, restriction digestion was done by adding 1 µL DpnI directly to the PCR product sample and incubating at 37 °C for 1 h. Transformation of a 45 µL aliquot of ultracompetent *Escherichia coli* XL10-Gold cells (provided with the kit) with 2 µL DpnI-digested sample was done as per the manufacturer's instructions. Transformed cells were plated on LB agar containing 100 µg/mL kanamycin and incubated at 37 °C for 24 h. Three colonies were selected and sub-cultured, plasmid DNA was extracted using the GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) and quantified using the NanoDrop™ One spectrophotometer (Thermo Fisher Scientific).

The V30ΔC1-d(22) mutant was generated by introducing a premature stop codon in place of the codon for canonical Rep Gln338. Site directed mutagenesis was conducted using Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher Scientific) and partially overlapping mutagenesis primers were designed following published recommendations (Laible and Boonrod, 2009; Liu and Naismith, 2008; Zheng et al., 2004) with amendments as described in Chapter 2, Section 2.2.3. Mutagenesis

primers were synthesised and RP-cartridge purified by Inqaba Biotech (Pretoria, SA). Mutagenesis PCR was set up as per manufacturer's instructions using 4 ng V30 template DNA with 1 % DMSO being included to inhibit ss formation. A two-step cycling protocol (72 °C annealing temperature) was followed with 20 cycles and an extension time of 520 s (45 s/kb). Upon completion, 5 µL sample was resolved on 0.7 % agarose to verify the presence of the target product at ~12 kb. Restriction digestion using DpnI (Thermo Fisher Scientific) was done by adding 1 µL enzyme directly to the remaining PCR product sample and incubating at 37 °C for 15 h followed by heat inactivation. Preparation and transformation of ultracompetent *E. coli* XL10-Gold cells was done as described in Chapter 2, Section 2.2.3. All mutant clones generated in this study were confirmed using PCR with Sanger sequencing. Sequence chromatograms for C1 deletion mutants are shown in A2 Fig. 3. Details of mutagenesis and sequencing primers used are indicated in Supplementary 3 (S3) Table 3.1.

3.2.2 Recombinant V30 RepC expression vector construction

The partial V30 C1 ORF nt sequence encoding canonical Rep aa 122-359 containing the helicase and C-terminal domains (RepC) was codon-optimised for expression in *E. coli* and subsequently cloned into pCold TF expression vector using BamHI and XbaI by Genscript Biotech (Piscataway, USA) to generate pCold TF-RepC. The predicted target protein product was termed His-TF-RepC. Transformation of NEB® 5-alpha Competent *E. coli* cells (New England Biolabs) was done according to the manufacturer's high efficiency transformation protocol. Transformed cells were plated on LB agar containing 100 µg/mL carbenicillin and incubated at 37 °C for 24 h. Three colonies were selected and sub-cultured, plasmid DNA was extracted and quantified.

The pCold I/TF and pMAL-c2X recombinant protein expression vectors were kindly provided by Dr IA Achilonu (University of the Witwatersrand, SA) and Prof S Chakraborty (Jawaharlal Nehru University, India) respectively. To subclone the codon-optimised RepC nt sequence into these two vectors, 1 µg aliquots of pCold TF-RepC, pCold I and pMAL-c2X were each subjected to double digestion with FastDigest BamHI and FastDigest XbaI (Thermo Fisher Scientific) with incubation at 37 °C for 1 h followed by heat inactivation. The products of digestion were resolved on 0.8 % agarose. Gel slices containing bands corresponding to the RepC insert, as well as

linearised pCold I and pMAL-c2X were excised using sterile scalpels and DNA was extracted using the Macherey-Nagel™ NucleoSpin™ Gel and PCR Clean-up Kit (Thermo Fisher Scientific). Extracted DNA was quantified, and ligation was conducted using T4 DNA Ligase (Thermo Fisher Scientific) with an insert:vector molar ratio of 3:1. Reactions were incubated at 16 °C for 16 h with subsequent heat inactivation. NEB® 5-alpha Competent *E. coli* cells (New England Biolabs) were transformed with ligation reactions according to the manufacturer's high efficiency transformation protocol. Transformed cells were plated on LB agar containing 100 µg/mL carbenicillin and incubated at 37 °C for 24 h. Three colonies were selected and sub-cultured, plasmid DNA was extracted and quantified. Restriction digestion was used to check for the presence of the RepC insert in the modified expression vectors termed pCold I-RepC and pMAL-c2X-RepC. The predicted target protein products of the modified expression vectors were designated His-RepC and MBP-RepC respectively.

Site directed mutagenesis of pCold I-RepC and pMAL-c2X-RepC was conducted using the QuikChange® II XL Site-Directed Mutagenesis Kit (Agilent Technologies) to generate modified expression vectors encoding 14 aa C-terminal deletion mutants of RepC. The resulting mutated expression vectors were designated pCold I-RepC_14 and pMAL-c2X-RepC_14. Premature stop codons were introduced in place of the codon for canonical Rep Gln346 of the RepC insert present in both expression vectors. Mutagenesis primers were designed using the QuikChange® Primer Design Program web server and synthesised/purified as before. The QuikChange® II XL Site-Directed Mutagenesis Kit (Agilent Technologies) was used for mutagenesis, with 10 ng of each template expression plasmid being used for separate mutagenesis PCR reactions. A three-step cycling protocol was followed with 18 cycles and an extension time of 330 s for pCold I-RepC template and 480 s for pMAL-c2X-RepC template (60 s/kb). DpnI digestion of PCR products and transformation of ultracompetent *E. coli* XL10-Gold cells was carried out as before (Section 3.2.1). Transformed cells were plated on LB agar containing 100 µg/mL carbenicillin and incubated at 37 °C for 24 h. Three colonies per mutant transformation were selected and sub-cultured, plasmid DNA was extracted and quantified. All modified recombinant protein expression vectors and mutated clones generated in this study were confirmed using PCR with Sanger sequencing (Inqaba Biotech, Pretoria). Sequence chromatograms for expression vector deletion mutants are shown in A2 Fig. 4. Details of mutagenesis and

sequencing primers used are indicated in S3 Table 3.1. Details of nt and predicted aa sequences of recombinantly expressed RepC proteins in S3 Fig. 3.1-3.3.

3.2.3 Amino acid sequence analyses and homology modelling

Iteron-like conserved repeated sequence elements (putative Rep-binding sites) present in the 5' end of the IR of DNA-A were manually identified for ToCSV V30, V22, exemplar isolate (AF261885), TYLCSV (X61153) and selected African mono- and bipartite begomoviruses. These iteron-like motifs were situated upstream of the canonical C1 start codon and on either side of the conserved TATA box. The organisation of the identified motifs was determined for each isolate and compared.

Following standard code, the ORF finder server (https://www.ncbi.nlm.nih.gov/orffinder/) was used to identify the C1 ORF encoding the Rep proteins of V30, V22 (described in Chapter 2, Section 2.2.1), TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. Functional domain coordinates were identified for all canonical Rep protein sequences by following the TYLCSV Rep domain boundary delineation described by Clérot and Bernardi (2006). An alignment of canonical Rep protein aa sequences and a separate alignment of Rep proteins found to contain putative N-terminal extensions was generated using the MUSCLE alignment program with the EMBL-EBI web server (Madeira et al., 2019). Begomovirus Rep protein alignments were further analysed using the MView v 1.63 alignment server (Brown et al., 1998). Rep protein ss was predicted using PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) (Buchan and Jones, 2019; Jones, 1999) and regions of disorder were predicted using the DISOPRED web server (Jones and Cozzetto, 2015). Sequence alignments were manually annotated, and details of all begomovirus isolates used are shown in S3 Table 3.2

Since no solved three-dimensional structures of ToCSV Rep protein exist in the Protein Data Bank (PDB), homology modelling was done using the Phyre2 (Kelley et al., 2015) and SWISS-MODEL (Waterhouse et al., 2018) web servers. The full V30 Rep aa sequence (aa 1-359) was submitted as the target sequence. Two 3D structures were obtained using the Phyre2 server: N-terminal domain (aa 1-121, monomer) and Rep helicase and C-terminal domain (aa 181-359, monomer). Two 3D

structures were obtained using the Phyre2 server: N-terminal domain (aa 4-121, monomer) and Rep partial helicase and C-terminal domain (aa 211-341, hexamer). The most significant template identified for the N-terminal domain of V30 Rep (aa 4-121) was the structure of TYLCSV Rep (aa 4-121) solved using NMR spectroscopy (PDB ID: 1L5I) sharing 81.4 % sequence identity. The most significant template identified for the partial C-terminal domains of V30 Rep (aa 211-341) was the hexameric structure of the helicase domain of BPV E1 helicase (aa 423-549) solved using X-ray crystallography (PDB ID: 2GXA) sharing 15.4 % sequence identity. All four homology models were subjected to structure quality assessment using the SWISS-MODEL web server. Homology models were visualised as semi-transparent surface and cartoon representations in 3D space using edu PyMOL (The PyMOL Molecular Graphics System, v 1.7.4 Schrödinger, LLC). Functional motifs were highlighted, and important residues were depicted using liquorice representation.

The MUSCLE alignment program was used to generate an aa sequence alignment of N-terminal domain functional motifs of V30 Rep, TYLCSV Rep (NP_620741), WDV Rep (AHZ63100), faba bean necrotic yellows virus (FBNYV) M-Rep (CAA72209), and porcine circovirus 2 (PCV2) Rep (AAQ94098). An alignment of helicase domain functional motifs of V30 Rep, TYLCSV Rep, WDV Rep, PCV2 Rep, AAV2 Rep (QDH44425), BPV E1 (NP_056739), and simian vacuolating virus 40 (SV40) Tag (NP_043127) was also produced. Functional motifs were mapped to partial solved structures of TYLCSV Rep (PDB ID: 1L5I), WDV Rep (PDB ID: 6Q1M), FBNYV M-Rep (PDB ID: 2HWT), PCV2 Rep (PDB ID: 2HW0 and 7LAR), AAV2 Rep (PDB ID: 7JSI), BPV E1 (PDB ID: 2GXA), and SV40 Tag (PDB ID 1SVM) using edu PyMOL. Protein topology diagrams were manually constructed with the Paint 3D program (Microsoft) for comparative analysis of functional motif positions and PyMOL-assigned three state ss (helix/sheet/loop) for all molecular structures analysed. In the case of hexameric structures, the consensus monomer PyMOL-assigned ss was depicted.

The biochemical properties of the predicted aa sequences of all recombinantly expressed RepC proteins before and after tag cleavage were determined using the ProtParam program (Gasteiger et al., 2005) with the ExPASy Server. Predicted XtalPred Crystallizability classification of selected recombinantly expressed RepC proteins after tag cleavage were determined using the random forest classifier (RF) method with the XtalPred web server (Jahandideh et al., 2014; Slabinski et al., 2007).

3.2.4 V30 C1 putative 5’ end transcript detection

To investigate the presence of an RNA transcript derived from the putative 5’ extension of the C1 ORF of V30, specific primers for RT-PCR targeting this region were designed using Primer3 (Kõressaar et al., 2018). Primer sequences were evaluated for the presence of ss using the Oligoanalyzer™ Tool (https://www.idtdna.com/oligoanalyzer). Total RNA extracted previously from V30-infected and mock-inoculated *N. benthamiana* leaf tissue (Chapter 2, Section 2.2.6) was used for first strand cDNA synthesis. The RevertAid™ First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) in accordance with the manufacturer’s protocol with inclusion of all prescribed negative controls including mock. PCR analysis was done using the DreamTaq™ PCR Master Mix (Thermo Fisher Scientific). A three-step cycling protocol (60 °C annealing temperature) was followed with 30 cycles and an extension time of 20 s. PCR products (target = 152 bp) were subsequently resolved on 2 % agarose. RT-PCR was repeated three times. For details of primers used refer to S3 Table 3.1.

3.2.5 In vivo infectivity trials of V30 C1 mutant infectious clones

Wild type *N. benthamiana* plants were prepared for agroinoculation with V30ΔC1-d(14) and V30ΔC1-d(22) and cultivated as described in Chapter 2, Section 3.2.4. Competent *Agrobacterium tumefaciens* C58C1 cells were transformed separately with the two C1 mutants and plant agroinoculation was carried out as per Chapter 2, Section 2.2.5. Infectivity trials of C1 mutants were run concurrently with infectivity trials conducted as part of Chapter 2 under identical conditions. Five plants were inoculated for each mutant and three trials were conducted. V30-inoculated plants, mock and non-inoculated controls were compared with C1 mutant-inoculated plants. All inoculated plants were observed daily for symptom development. C1 mutant-infected plants were observed for up to 60 dpi.

For DNA extraction, two leaf discs (d = 12.7 mm) were collected from expanded leaves below the meristem of plants inoculated with V30ΔC1-d(14) at 12, 20, and 28 dpi. Leaf tissue was flash frozen in liquid nitrogen and DNA was extracted following the method described in Chapter 2, Section 2.2.6. Viral load in V30ΔC1-d(14)-infected plants was

determined at 12, 20 and 28 dpi using real-time qPCR following the protocol described in Chapter 2, Section 2.2.7 and compared with viral load in V30-infected plants at identical timepoints. Primers used for qPCR are shown in S3 Table 3.1.

3.2.6 Recombinant RepC expression

Upon confirmation of recombinant RepC expression construct sequences with restriction digestion and/or PCR with Sanger sequencing, protein expression trials in *E. coli* were conducted. The three strains of *E. coli* used in expression trials were BL21(DE3), BL21(DE3) pLysS, and T7 Express. Competent cells were prepared by following the protocol by Green and Sambrook (2020). For transformation, competent cell aliquots (50 µL) were thawed and ~50 ng plasmid DNA was added to each aliquot which was subsequently kept on ice for 30 min. After heat shock transformation (60 s, 42 °C water bath) cells were immediately transferred onto ice for two minutes. Thereafter, 450 µL pre-warmed SOC medium was added to each aliquot and cells were incubated at 37 °C, with shaking (200 rpm) for 1 h. Transformed cells were plated on either SOB (with 10 mM $MgSO_4$) or LB agar containing carbenicillin (100 µg/mL) and incubated at 37 °C overnight. For plating of transformed BL21(DE3) pLysS, chloramphenicol (34 µg/mL) was also included in the medium. Preliminary transformations revealed that BL21(DE3) and BL21(DE3) pLysS transformed with pCold I-RepC was unable to grow adequately in liquid culture. As a result, the T7 Express strain was used for His-RepC and MBP-RepC protein expression since this strain grew normally when transformed with pCold I-RepC and pMAL-c2X-RepC.

To conduct subsequent RepC expression trials, specific parameters were applied. For pCold vectors containing RepC, the IPTG concentrations (0.1-1.0 mM), induction temperature (15 °C) and expression time (24 h) used were as recommended by the vector manufacturer (Takara Bio). Culture media used for pCold vector mediated RepC expression included SOB (with 10 mM $MgSO_4$) or LB. For pMAL-c2X-RepC, IPTG concentrations tested ranged from 0.1-0.4 mM. An expression temperature of 18 °C for 16 h was used as this was previously reported for pMAL-c2X expression of other begomovirus Rep proteins (Clérot and Bernardi 2006; George et al., 2014). For liquid culture of cells transformed with pMAL-c2X-RepC, rich medium (RM: 10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl, 2 g/L glucose) was used. The inclusion of

glucose ensures repression of *E. coli* chromosomal genes which encode enzymes that result in degradation of amylose that negatively impacts downstream purification with amylose resin.

After transformation, single colonies were selected and subcultured in 5 mL liquid broth containing appropriate antibiotics at 37 °C with shaking at 200 rpm. Preliminary expression trials revealed that the use of overnight cultures did not result in consistent downstream protein expression. It was observed that 5 mL “outgrowth” cultures that reached OD_{600} 1.0 in a period of ≤7 h were more likely to result in successful protein expression. This time limit was consequently applied in selecting suitable transformants for downstream protein expression. In addition, glycerol stocks of transformed cells did not yield reproducible RepC protein expression levels and fresh transformation was necessary prior to subsequent protein expression. Outgrowth cultures (2 mL) were centrifuged (6000xg, 2 min, RT) and resuspended in 1 mL fresh antibiotic-free medium. Baffled flasks (with membrane caps for improved aeration) containing 100 mL liquid medium supplemented with antibiotics and 0.005 % Antifoam 204 (Merck) were each inoculated with 1 mL resuspended outgrowth culture. Cultures were incubated at 37 °C with shaking (200 rpm) until OD_{600} ~0.6 for cultures containing pCold vector with RepC and OD_{600} ~0.8 for cultures containing pMAL-c2X-RepC. Prior to induction with IPTG, cultures containing pCold vector with RepC were chilled in an ice bath for 30 min. IPTG was added at the appropriate concentrations and flasks were incubated with shaking at the abovementioned temperatures for the specified time periods. After expression, cultures were centrifuged (6000xg, 15 min, 4 °C), wet cell pellet masses were determined, and pellets were subsequently stored at -80 °C until use.

For extraction of His-TF-RepC and His-RepC, lysis buffer was used at 5 mL/g cell pellet and contained 50 mM NaH_2PO_4 (pH 8.0), 300 mM NaCl, 5 mM imidazole, 1 mM PMSF and 10 % glycerol (Buffer A1). The lysis buffer used to extract MBP-RepC contained 50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 5 mM $MgCl_2$, 2 mM β-mercaptoethanol, 1 mM PMSF and 10 % glycerol (Buffer A2). Cell pellets were thawed and lysozyme was added to a final concentration of 1 mg/mL prior to resuspension. After gently mixing to resuspend the pellets, samples were mixed on ice for 30 min. Thereafter cell disruption on ice using sonication (40 % amplitude, 30 s on/30 s off, 5 cycles) was done and 50 µL cell lysate samples were collected and kept on ice. The

remaining sample volume was centrifuged (20 000xg, 30 min, 4 °C) to separate soluble and insoluble fractions. Sample fractions were analysed with SDS-PAGE using the TGX Stain-Free™ FastCast™ Acrylamide Kit (Bio-Rad Laboratories). All extracted soluble fractions confirmed to contain the target protein were kept on ice at 4 °C until purification. Samples were purified immediately, and purification was conducted at 4 °C (where possible) using chilled filter-sterilised and degassed buffers.

3.2.7 Recombinant RepC purification

Soluble fractions confirmed to contain the target His-TF-RepC protein were pooled and subjected to immobilised metal affinity chromatography (IMAC) ÄKTA purification using a 5 mL HisTrap FF column containing immobilised Ni^{2+} resin (GE Healthcare). The fractions were loaded onto the column which was equilibrated with buffer A1. The column was subjected to three sequential washes: buffer A1, buffer A1 containing 0.5 % TWEEN 20 (detergent wash), and buffer A1 containing 1 M NaCl (high salt wash). Elution was conducted using buffer A1 containing 500 mM imidazole. Flowthrough, wash and elution fractions were collected and analysed using SDS-PAGE. Upon target confirmation, an elution sample (~0.7 mg) was subjected to dialysis (1^{st}: 6 h, 2^{nd}: 10 h, 4°C) against buffer B1 containing 50 mM Tris-HCl (pH 8.1), 100 mM NaCl, 5 mM $CaCl_2$, 5 mM $MgCl_2$ and 10 % glycerol. The sample was used in a proteolytic cleavage pilot experiment with 1 U thrombin (Roche) at 4 °C for up to 16 h with samples being collected for SDS-PAGE analysis at 5 min, 15 min, 30 min, hourly until 7 h, and at 16 h. Once the optimal cleavage time was determined (~ 6 h) the remaining IMAC elution fraction was dialysed against buffer B1, treated with thrombin, and then applied to a heparin-agarose affinity chromatography column to remove thrombin. Eluted protein was dialysed (16 h, 4 °C) against buffer A1 and subjected to a second IMAC (Ni^{2+} resin) ÄKTA purification. Flow through and elution fractions were separately pooled, dispensed into Amicon® Ultra-15 Centrifugal Filters (Merck) and centrifuged (5000 g, 10 min intervals, 4 °C) until a 1 mL concentrated volume was obtained for each. Samples were then analysed using SDS-PAGE.

His-RepC obtained in the soluble fraction was subjected to IMAC gravity flow purification using an Econo-Column® glass chromatography column 1.5 x 10 cm (Bio-Rad Laboratories) with a 3 mL resin bed composed of IMAC Sepharose® 6 Fast Flow

resin (Merck) charged with Ni^{2+}. Pooled soluble fraction was filtered and applied to the column after equilibration with buffer A1. The column was then washed with buffer A1 and bound protein was eluted using buffer A1 containing 500 mM imidazole. Flowthrough, wash and elution fractions were collected and analysed using SDS-PAGE.

Purification of His-RepC present in extracted soluble fraction using ammonium sulfate [AS: $(NH_4)_2SO_4$] precipitation was conducted as an alternative method to IMAC. Varying amounts of solid $(NH_4)_2SO_4$ was added to 2 mL soluble fraction aliquots to obtain a range of percentage saturation (10, 20, 25, 35 and 45 %). The mass of $(NH_4)_2SO_4$ required to obtain each desired concentration was determined according to Duong-Ly and Gabelli (2014a). Samples were mixed and incubated with gentle rotation at 4 °C for 1 h. Samples were then centrifuged (15 000xg, 5 min, 4 °C) and pellets (if obtained) were separated from supernatant and resuspended in buffer A1. Supernatant and pellet samples were analysed using SDS-PAGE.

His-RepC purification was also attempted using gel filtration chromatography (gravity flow) using an Econo-Column® glass chromatography column 1.5 x 50 cm (Bio-Rad Laboratories) containing sepharose 6B agarose (Merck) prepared by Dr IA Achilonu (University of the Witwatersrand, SA). His-RepC was expressed and extracted following protocols described in Section 3.2.6 using buffer A3 containing 50 mM Tris-HCl (pH 8.0), 1 M NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1 mM PMSF, 20 % glycerol, 1 % TWEEN 20 and 0.1 % NaN_3. Soluble fractions containing the target protein were pooled, dispensed into dialysis tubing, and concentrated using solid sucrose until sample volume was reduced by ~60 %. The concentrated soluble fraction was applied to the gel filtration column equilibrated with buffer A3. Elution fractions were collected and A280 values were immediately measured using the NanoDrop™ One spectrophotometer (Thermo Fisher Scientific) with the 1 Abs=1 mg/mL “sample type” setting selected for protein mixtures. Fractions were collected until A280 values were consistently near zero. All elution fractions collected were analysed using SDS-PAGE.

MBP-RepC expressed in the soluble fraction was subjected to amylose affinity chromatography (gravity flow) using an Econo-Column® glass chromatography column 1.5 x 10 cm (Bio-Rad Laboratories) with a 1 mL bed consisting of amylose resin (New England Biolabs). Pooled soluble fraction was diluted 1:6 with buffer A2

and applied to the column equilibrated with buffer A2. The column was washed with buffer A2, and bound protein was eluted using buffer A1 containing 10 mM maltose. Flowthrough, wash, and elution fractions were collected and analysed using SDS-PAGE. After confirming the presence of the target protein in the elution fraction, a sample (~ 0.01 mg) was aliquoted and $CaCl_2$ was added to a final concentration of 2 mM. A proteolytic cleavage pilot experiment was conducted by adding 0.2 µg factor Xa (New England Biolabs) and incubating at 23 °C for up to 24 h with samples being collected for SDS-PAGE analysis at 5 min, 30 min, 1 h, 2 h, 4 h, 8 h, 10 h, 12 h, 16 h, and at 24 h. Once the optimal cleavage time was determined (24 h), the remaining elution fractions containing MBP-RepC were subjected to factor Xa cleavage. Protein was then dialysed (1^{st}: 3 h, 2^{nd}: 3 h, 3^{rd}: 10 h, 4°C) against buffer B2 containing 20 mM Tris-HCl (pH 8.0), 1 mM DTT, and 10 % glycerol. Dialysed sample was subjected to ion-exchange chromatography with the NGC Quest 10 Plus Chromatography System (Bio-Rad Laboratories) using a 5 mL HiTrap DEAE FF anion exchanger column (GE Healthcare). The sample was centrifuged and loaded onto the column equilibrated with buffer B2. The column was washed with buffer B2 followed by a continuous NaCl gradient elution step with buffer B2 containing 1 M NaCl. A total gradient volume of 125 mL (0-100 % elution buffer, 5 mL/min) was applied. Flowthrough, wash, and elution fractions were collected and analysed using SDS-PAGE.

3.2.8 Comparative miniaturised recombinant RepC and RepC_14 expression

To determine the effect of the 14 aa C-terminal deletion mutation on the expression of RepC, a comparative protein expression experiment was designed by applying principles of miniaturisation and parallelisation. Two unmutated expression vectors (pCold I-RepC and pMAL-c2X-RepC) and their corresponding mutant vectors (pCold I-RepC_14 and pMAL-c2X-RepC_14) were used. Generation of these two mutants is described in Section 3.2.2. Aliquots (100 µL) of competent *E. coli* T7 Express prepared previously were separately transformed with ~100 ng each vector using heat shock (90 s, 42 ° C water bath). Cells were kept on ice for 5 min and thereafter 800 µL pre-warmed SOC was added. Cells were incubated at 37 °C, with shaking (200 rpm) for 1.5 h.

The induction temperatures, expression times, and IPTG concentrations applied for small scale expression (1 mL) using the four vectors were identical to those described for 100 mL expression with the unmutated vectors in Section 3.2.6. After growth in SOC medium, 50 uL volumes of each transformation were used to inoculate 3 x 5 mL LB containing carbenicillin (80 µg/mL) and 0.005 % Antifoam 204 (Merck). Cultures were incubated at 37 °C with shaking (200 rpm) until OD_{600} ~0.6 for cultures containing pCold I with RepC/RepC_14 and OD_{600} ~0.8 for cultures containing pMAL-c2X with RepC/RepC_14. Thereafter, 1 mL aliquots were dispensed in triplicate from each of the 3 x 5 mL cultures into sterile McCartney bottles. Bottles with cultures containing pCold I with RepC/RepC_14 were chilled in an ice bath for 5 min and then incubated at 15 ° C for 15 min without shaking. Bottles with cultures containing pMAL-c2X with RepC/RepC_14 were incubated at 18 °C for 15 min without shaking. The bottles were placed in a holding box and the appropriate volume of IPTG was added to each 1 mL culture in triplicate for induction following the prescribed temperatures/times with shaking at 150 rpm. Each box contained a total of three uninduced cultures, nine IPTG-induced cultures and one non-transformed culture. After expression, cultures were centrifuged (6000xg, 15 min, 4 °C), wet cell pellet masses were determined, and pellets were stored at -80 °C.

Cell pellets were each resuspended in 400 µL lysis buffer A4 containing 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 5 mM MgCl2, 2 mM β-mercaptoethanol, 1 mM PMSF, and 10 % glycerol. After pellet resuspension, cell disruption on ice using sonication (40 % amplitude, 20 s on/30 s off, 3 cycles) was done and 10 µL cell lysate samples were collected and kept on ice. The remaining sample volume was centrifuged (20 000xg, 30 min, 4 °C) to separate soluble and insoluble fractions. Fraction samples were analysed using SDS-PAGE.

3.3 Results

3.3.1 In silico analyses of African begomovirus Rep proteins reveals conserved functional motifs typical of other viral Rep proteins.

Although the primary aim of this study was to investigate the aa sequences of several African begomovirus Rep proteins, an analysis of the 5’ IR region of the corresponding DNA-A sequences of these viruses was done concurrently to identify TATA-associated

iteron-like sequence elements which may function as Rep-binding sites. These iteron-like motifs are bound by Rep in a sequence-specific manner. Comparative analysis of iteron-like motifs with their cognate Rep protein sequences may improve the understanding of the role of this DNA-protein interaction in disease development. Close inspection of the 5' end of the IR of ToCSV variant and selected African begomovirus sequences showed that iteron-like sequences display diverse sequence identity as well as arrangement (Fig. 3.1). The sequence of TYLCSV, a European begomovirus used as the outgroup in phylogenetic analysis reported in Chapter 2, was also included in sequence analysis.

All isolates contained two to three iteron-like sequences upstream of the conserved TATA element. The iteron motif immediately upstream of the TATA element was designated as the "primary" direct repeat with which all other iterons present were compared. Iterons containing a maximum of two nt mismatches were included. Nearly all isolate sequences analysed were shown to contain a single inverted repeat iteron immediately downstream of the TATA element. The sequence of the primary direct repeat of both ToCSV V30 and V22 (TCGGTACCCA) was identical to that of TbLCZV and differed by one nt with that of tomato leaf curl Sudan virus (ToLCSDV/Sha) and tomato leaf curl Comoros virus (ToLCKMV). The iteron organisation (5'-3': inverted, direct, inverted) was identical for ToCSV variants and these three other begomovirus isolates.

A multiple aa sequence alignment of the canonical Rep N-terminal domain (Fig. 3.2) revealed that the IRD is located at the N-terminal end of Rep encompassing the first ten residues. The IRD is a subdomain of the N-terminal domain and is believed to be responsible for the interaction with iteron sequences in the IR. The sequence of the IRD was not found to be highly conserved (even at 70 % consensus) and only the N-terminal methionine and a phenylalanine at alignment position 7 were conserved at 100 % consensus (Fig. 3.2A). The IRD sequence of ToCSV was identical to that of TbLCZV and ToLCSDV/Sha. The IRD of PepYVMLV and ToLCDiV each differed from the ToCSV IRD by one residue. The IRD of all analysed begomovirus Rep proteins were predicted to contain an N-terminal region of disorder 1-4 aa in length with 16 of the proteins (including ToCSV Rep) predicted to contain regions of disorder with protein binding potential (Fig. 3.2B).

Fig. 3.1 Organisation of iteron-like conserved repeated sequence elements present in the 5' end of the intergenic region (IR) of DNA-A of ToCSV V30, V22, exemplar isolate (AF261885), TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. Virion-sense italicised CAT sequence corresponding to canonical Rep start codon. Orientations of iteron-like motifs shown using arrows. Violet arrows indicate iterons conserved with the first iteron sequence upstream of the TATA element in bold, grey arrows indicate iterons with mismatches when compared with the first iteron upstream of the TATA element. Ellipses indicate spacer sequence with nt number shown. Details of all isolate sequences used in S3 Table 3.2.

Four motifs implicated in the binding, cleavage and replication of viral DNA are present in the N-terminal domain and their positions are highlighted in Fig. 3.2A. In the multiple sequence alignment, Motif I had a 100 % consensus sequence of FℓT(F/Y)P, where ℓ is an aliphatic residue. Motif II had a 100 % consensus of GXXH(I/L)HhLh, where X is unconserved, and h is a hydrophobic residue. The GRS motif contained a conserved N-terminal 100 % consensus sequence RhFD and a C-terminal 100 % consensus sequence of FHPNhQGAKS both separated by a series of small and polar residues. Motif III had a 100 % consensus sequence of Yℓ(D/E)K(D/G). The N-terminal domain was predicted to contain four consensus β-strands and two helices (confidence ≥5) (Fig. 3.2B). Two additional short consensus β-strands were predicted with low confidence (<5). Three C-terminal IRD residues form part of a β-strand (confidence <5). The first three residues of Motif I are located at the C-terminal end of the second consensus β-strand. Most of the motif II residues were predicted to be part of the fourth consensus β-strand. The region spanning the GRS lacks consensus β-strands or helices and was predicted to contain short β-strands or helices in some Rep

proteins with low confidence (<5). The second consensus helix contains the first two residues of Motif III (confidence ≥5).

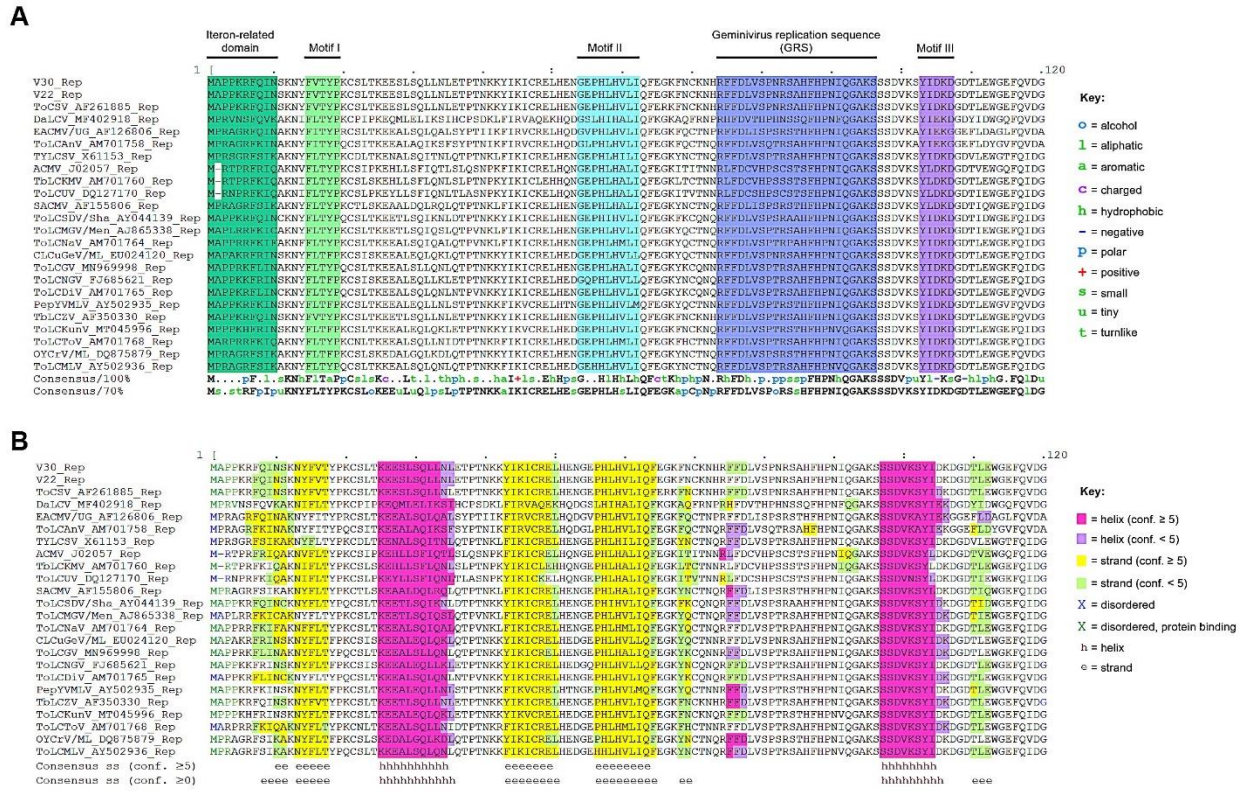


Fig. 3.2 Rep N-terminal domain multiple sequence alignment for ToCSV V30, V22, TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. **A**, MUSCLE aa sequence alignment with 100 % and 70 % aa sequence consensus shown. Consensus residues capitalised and coloured lowercase letter key indicates consensus residue type. Functional sequence motifs highlighted and annotated. **B**, PSIPRED secondary structure (ss) and DISOPRED protein disorder prediction, with ss elements highlighted and disordered residues coloured according to colour key. Consensus ss elements at confidence (conf.) levels ≥0 and ≥5 indicated. Rep sequence obtained with ORF finder using the DNA-A sequence as the input. DNA-A sequences for all isolates used in S3 Table 3.2.

The N-terminal domain of V30 Rep was modelled with Phyre2 (Fig. 3.3A) and SWISS-MODEL (Fig. 3.3B). The homology models are herein referred to as “RepN-P” and “RepN-SM” respectively. The template used to model the monomeric homology models was the structure of TYLCSV Rep (aa 4-121, PDB ID: 1L5I) which shares 81.4 % sequence identity with V30 Rep N-terminal domain. Since the solved TYLCSV structure does not include the first three residues of Rep, these were modelled *ab initio* for V30 Rep by Phyre2 only. PyMOL visualisation revealed that both RepN-P and RepN-SM were shown to contain an antiparallel β-sheet comprised of five β-strands (β1-β5), and two helices (h1 and h2). The first helix was situated underneath the “core”

β-sheet whilst h2 was situated above it. RepN-P contained an additional short helical element positioned underneath β4. RepN-SM lacked this additional helix and instead contained two short β-strands forming a small antiparallel β-sheet situated underneath β4 of the large core β-sheet. Interestingly, two C-terminal residues of the IRD were mapped to the first short β-strand of the RepN-SM structure.

The surface representation of the two models (Fig. 3.3) revealed a probable catalytic site where residues belonging to motif I, II, III and the GRS were positioned close to each other. Residues belonging to motif I were mapped to β1, motif II to β3 and motif III to the C-terminal end of h2. The GRS motif was mapped to a long loop structure and four residues in the C-terminal half of the GRS formed β4. The positions of three important catalytic residues were shown using liquorice representation. Tyr103 of motif III and His57 and His59 of motif II were more closely positioned to each other in RepN-SM (Fig. 3.3B) than RepN-P (Fig. 3.3A).

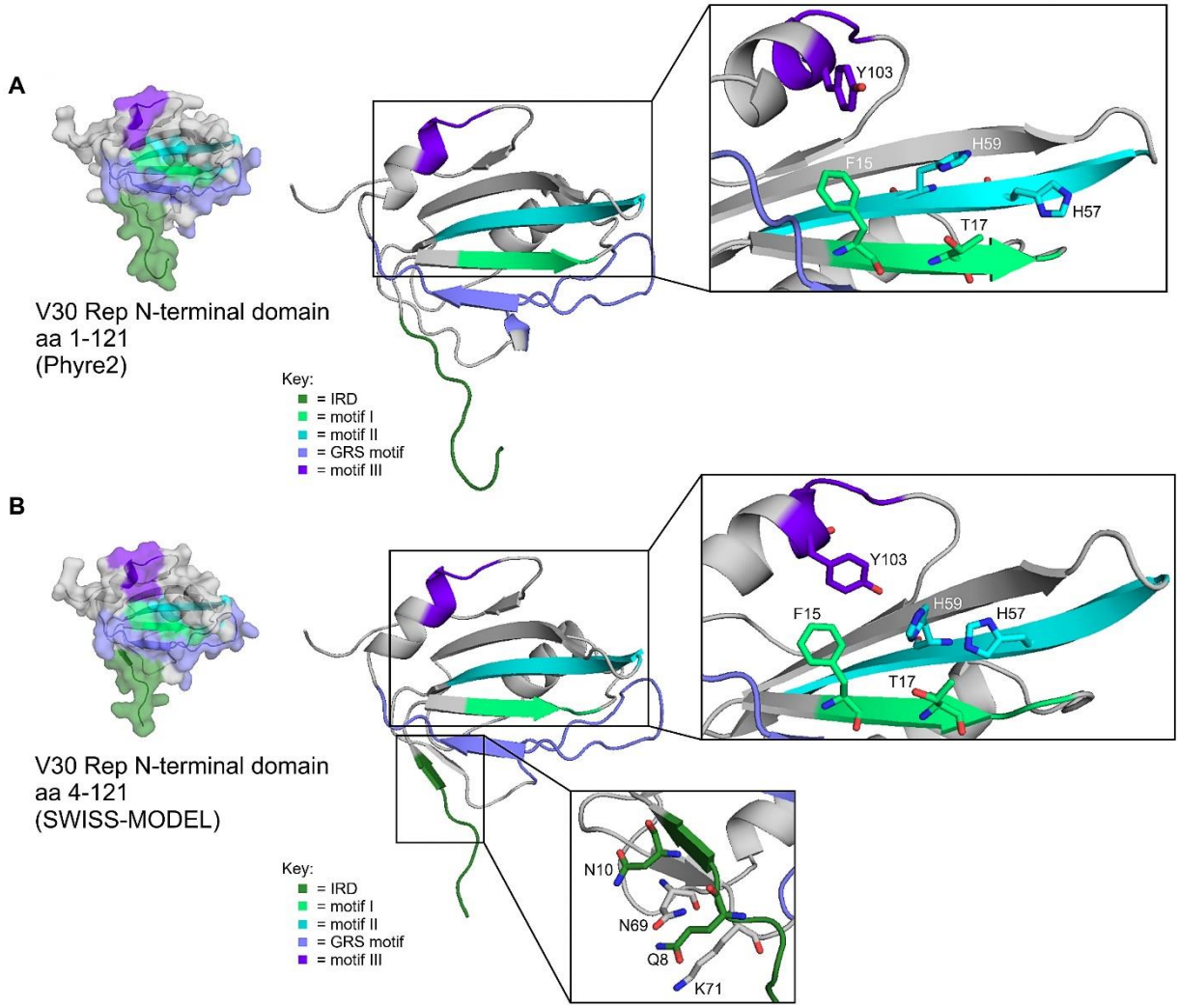


Fig. 3.3 Semi-transparent surface and cartoon representation of ToCSV V30 Rep N-terminal domain homology model monomers generated using **A**, Phyre2 and **B**, SWISS-MODEL. Positions of functional motifs highlighted according to colour key with important IRD and IRD-associated, motif I, motif II and motif III residues depicted using liquorice representation. SWISS-MODEL structure quality assessment outputs shown in S3 Fig. 3.4 and S3 Fig. 3.5.

Quality assessment using SWISS-MODEL was done (S3 Fig. 3.4 and S3 Fig. 3.5) and the results revealed that RepN-SM scored better than RepN-P. The Ramachandran favoured/outlier percentages were 92.2/0.0 and 84.0/7.6 for RepN-SM and RepN-P respectively. The QMEAN Z-scores obtained were more favourable for RepN-SM than RepN-P. The QMEANDisCo Global scores obtained were 0.80 ± 0.08 for RepN-SM and 0.57 ± 0.08 for RepN-P. Taken together, these assessment outputs indicated that RepN-SM was of higher quality than RepN-P.

Sequence comparison of the four N-terminal domain functional motifs of V30 Rep with those present in the Rep protein sequences of other ssDNA viruses (Fig. 3.4A) was done to identify conserved residues present in each motif.

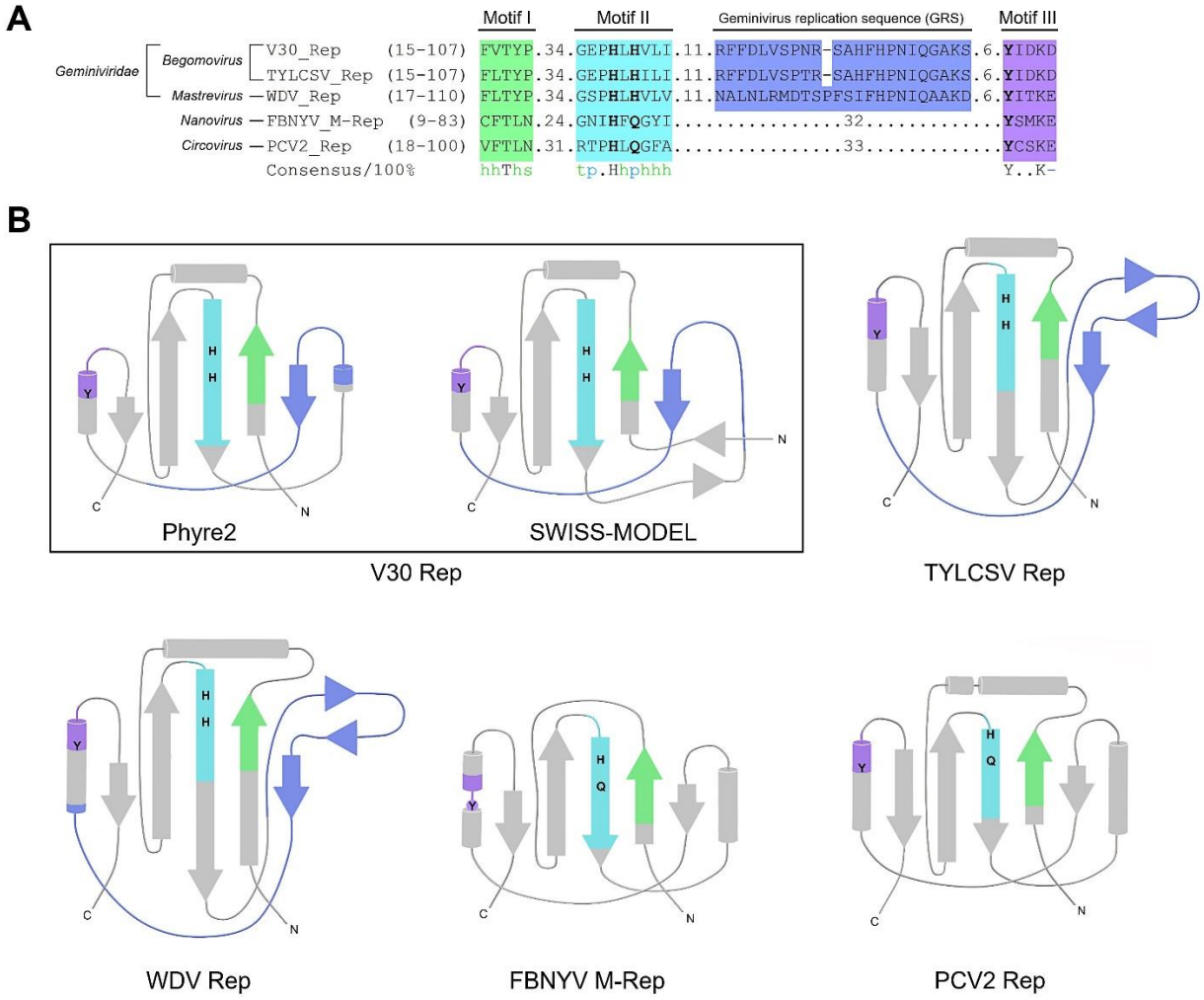


Fig. 3.4 Comparative analysis of selected ssDNA virus Rep N-terminal domains. **A**, MUSCLE aa sequence alignment of N-terminal domain functional motifs of ToCSV V30 Rep, TYLCSV Rep (NP_620741), WDV Rep (AHZ63100), FBNYV M-Rep (CAA72209), and PCV2 Rep (AAQ94098). Motifs highlighted and annotated, with important residues in bold. Consensus residues capitalised, consensus hydrophobic (h), negatively charged (-), polar (p), small (s) and turnlike (t) residues indicated with coloured lowercase letters. Hyphens indicate alignment-generated gaps, ellipses in alignment indicate spacer sequence with aa number shown, protein sequence aa positions in brackets. **B**, Structural topology of V30 Rep N-terminal domain homology model monomers generated using Phyre2 and SWISS-MODEL; and solved monomeric structures of TYLCSV Rep (PDB ID: 1L5I), WDV Rep (PDB ID: 6Q1M), FBNYV M-Rep (PDB ID: 2HWT), and PCV2 Rep (PDB ID: 2HW0). PyMOL-assigned "sheet" shown as arrows, "helix" shown as cylinders, "loop" shown as thin tubes. Position of motif I in green, motif II in cyan, geminivirus replication sequence in navy blue, motif III in violet. Positions of important motif residues indicated using bold letters. Details of PyMOL-assigned ss shown in S3 Fig. 3.6.

Motif I contained a consensus central threonine residue flanked on either side by a hydrophobic residue. Motif II contained a consensus Hh(H/Q)hhh with h being a hydrophobic residue. The third residue of this consensus sequence was a histidine in Rep of V30, TYLCSV and the related geminivirus WDV (genus *Mastrevirus*). The reside at this position was glutamine in the master Rep (M-Rep) of FBNYV, a plant-infecting nanovirus, as well as the Rep protein of PCV2 which is a pig-infecting circovirus. The catalytic tyrosine of motif III was conserved among all five proteins and the motif had a consensus sequence of YXXK(D/E), where X is an unconserved residue. The GRS motif has only been found in geminivirus Rep proteins and the C-terminal half of the motif in WDV Rep was highly conserved with that of V30 and TYLCSV. The GRS N-terminal arginine conserved amongst African begomovirus Rep proteins (Fig. 3.2A) was absent in WDV Rep (Fig. 3.4A).

Topology diagrams were constructed for RepN-P, RepN-SM, and the solved structures of the N-terminal domains of other ssDNA virus Rep proteins to compare the ss features present (Fig. 3.4B). PyMOL-assigned three state ss (helix/sheet/loop) was applied when constructing the diagrams (S3 Fig. 3.6). All structures were shown to contain the core antiparallel β-sheet composed of five β-strands surrounded by two to four helices. The relative positions of motif I, motif II, and motif III were highly analogous, and the location of key residues was consistent. The GRS motif of TYLCSV (PDB ID: 1L5I) and WDV (PDB ID: 6Q1M) Rep proteins was unusual in that both contained a β-hairpin loop structure just upstream of β4 that was absent in the V30 RepN homology models. RepN-SM was the only structure which contained a small antiparallel β-sheet situated underneath the core β-sheet.

Multiple sequence alignment of the central Rep oligomerisation domain of canonical V30 Rep (aa 121-180) with other begomovirus Rep sequences (S3 Fig. 3.7) revealed that this region was not highly conserved at 100 % consensus. The begomovirus oligomerisation domain was predicted to contain several long helices (S3 Fig. 3.7). A region of interest spanning the first nine residues of this domain was identified. This region contained two conserved positively charged arginine residues, as well as predicted regions of disorder for 13 out of 24 Rep proteins analysed. The Rep proteins of V30, V22, TbLCZV, and EACMV/UG were predicted to contain disordered residues with protein binding potential. Conserved functional motifs have not been described for the oligomerisation domain and this central region of begomovirus Rep shares no

significant sequence identity with central domains of other SF3 helicases. Consequently, homology modelling of this domain was not feasible.

The helicase domain of canonical V30 Rep (aa 181-330) was aligned with that of selected African begomovirus Rep proteins revealing a highly conserved core sequence at 100 % consensus level, spanning four functional motifs involved in ATP hydrolysis and helicase activity (Fig. 3.5A).

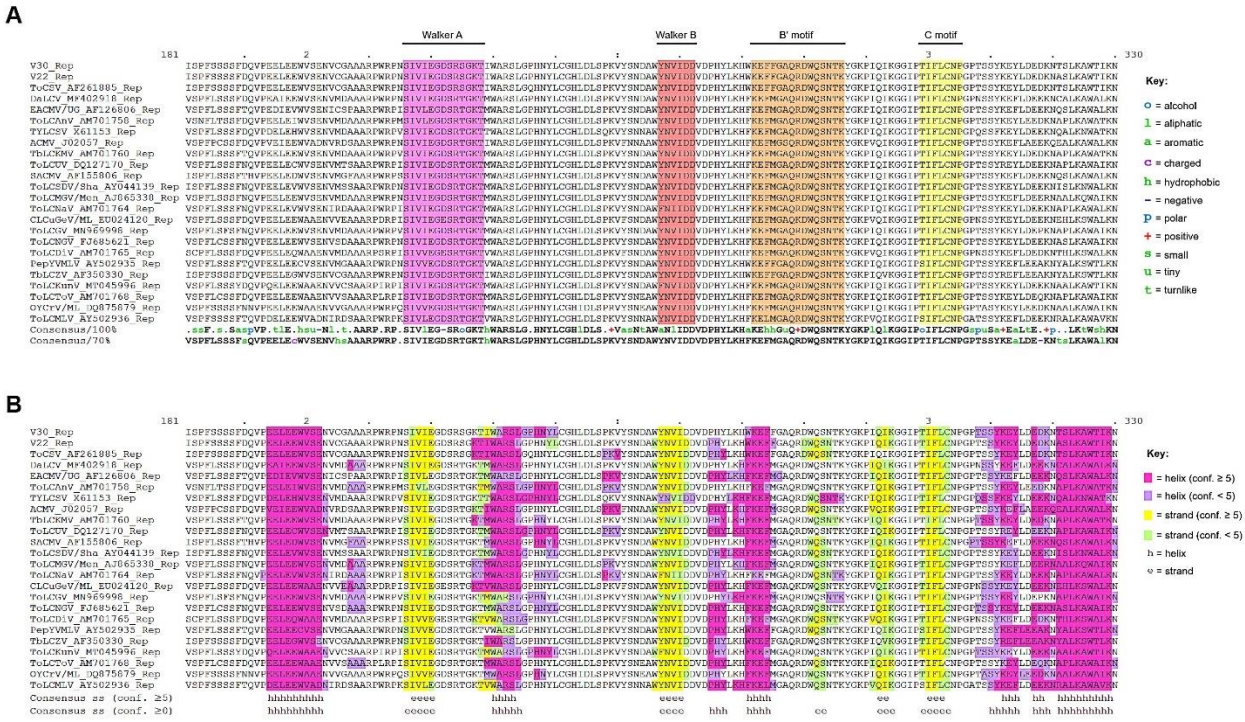


Fig. 3.5 Rep helicase domain multiple sequence alignment for ToCSV V30, V22, TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. **A**, MUSCLE aa sequence alignment with 100 % and 70 % aa sequence consensus shown. Consensus residues capitalised and coloured lowercase letter key indicates consensus residue type. Functional sequence motifs highlighted and annotated. **B**, PSIPRED secondary structure (ss) prediction, with ss elements highlighted according to colour key. Consensus ss elements at confidence (conf.) levels ≥0 and ≥5 indicated. Rep sequence obtained with ORF finder using the DNA-A sequence as the input. DNA-A sequences for all isolates used in S3 Table 3.2.

At 100 % consensus level the WA motif had a sequence of SIVℓEG(D/E)SR(S/T)GKT, WB motif had a sequence of (F/Y)N(I/V)IDD, B' motif had a sequence of KEhhG(A/S)Q(K/R)DWQSNTK, and the C motif had a sequence of (S/T)IFLCNP. The region spanning these motifs was predicted to contain four consensus β-strands and two short helices (confidence ≥5) (Fig. 3.5B). An additional short helix and short strand were predicted with low confidence (<5). Two helices were predicted outside of the

core domain, one in the N-terminal region and the other in the C-terminal region. Within the core region, WA residues at motif positions two to five were predicted to form the first β-strand. The first four residues of WB formed the second β-strand. The second helix contained the first three residues of the B' motif. A short β-strand was predicted to exist at B' motif residue positions eleven and twelve with low confidence (<5). The fourth consensus β-strand was predicted to be composed of C motif residues two to four.

The C-terminal domain (canonical V30 Rep aa 331-359) is the shortest Rep protein domain and its function is unknown. Multiple sequence alignment of V30 Rep C-terminal domain with other begomovirus Rep sequences (Fig. 3.6) showed that the sequence and length of this domain was highly variable. The two defining features of this domain included the presence of a consensus β-strand containing a conserved phenylalanine, two conserved residues at positions 339 (P) and 340 (L), as well as a C-terminal region 10-20 aa in length predicted to be disordered with protein-binding potential.

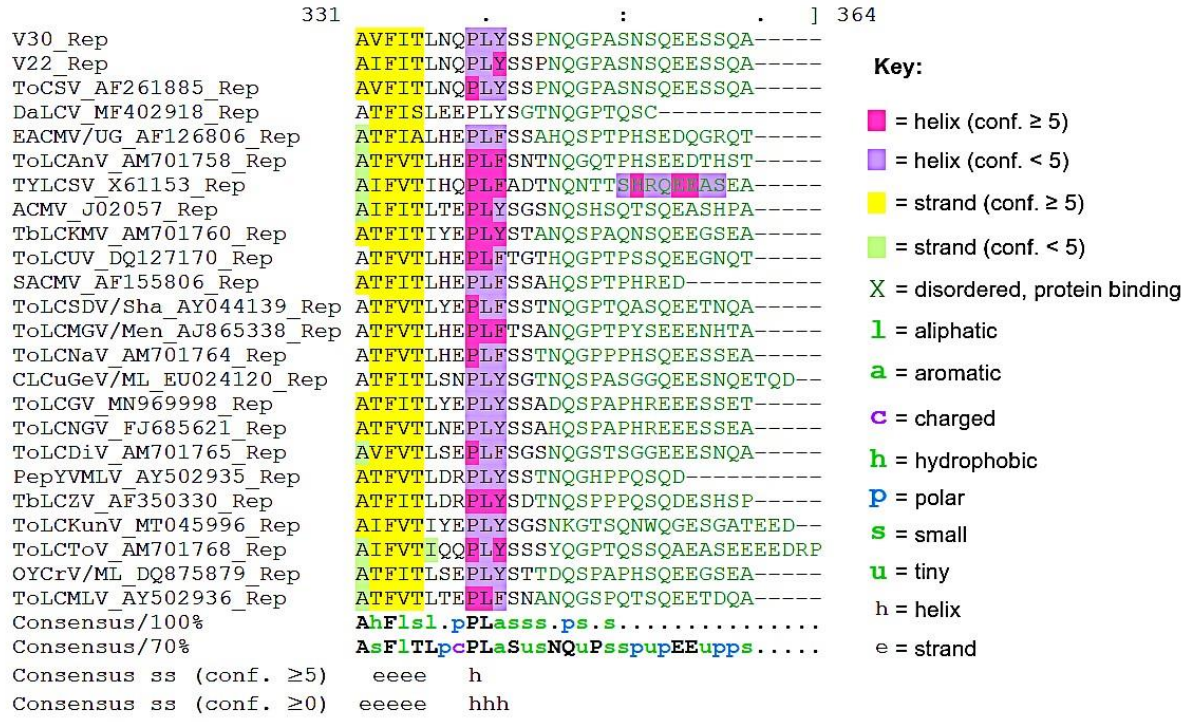


Fig. 3.6 Rep C-terminal domain multiple sequence alignment for ToCSV V30, V22, TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. MUSCLE aa sequence alignment with 100 % and 70 % aa sequence consensus shown. Consensus residues capitalised and coloured lowercase letter key indicates consensus residue type. PSIPRED secondary structure (ss) and DISOPRED protein disorder prediction, with ss elements highlighted and disordered residues coloured according to colour

key. Consensus ss elements at confidence (conf.) levels ≥0 and ≥5 indicated. Rep sequence obtained with ORF finder using the DNA-A sequence as the input. DNA-A sequences for all isolates used in S3 Table 3.2.

The full helicase and C-terminal domains of V30 Rep (aa 121-359) were modelled using Phyre2 (Fig. 3.7) and this monomeric structure is herein referred to as “RepC-P”. Residues corresponding to canonical Rep aa 121-129 and aa 342-359 were modelled *ab initio*. The partial helicase (aa 211-330) and C-terminal (aa 331-341) V30 Rep domains were also modelled using SWISS-MODEL (Fig. 3.8) with the resulting hexameric structure being termed “RepC-SM”. The common template used to produce the two models was the hexameric structure of the helicase domain of BPV E1 helicase (aa 423-549, PDB ID: 2GXA) which shares only 15.4 % sequence identity with the partial V30 Rep helicase and C-terminal domains (aa 211-341).

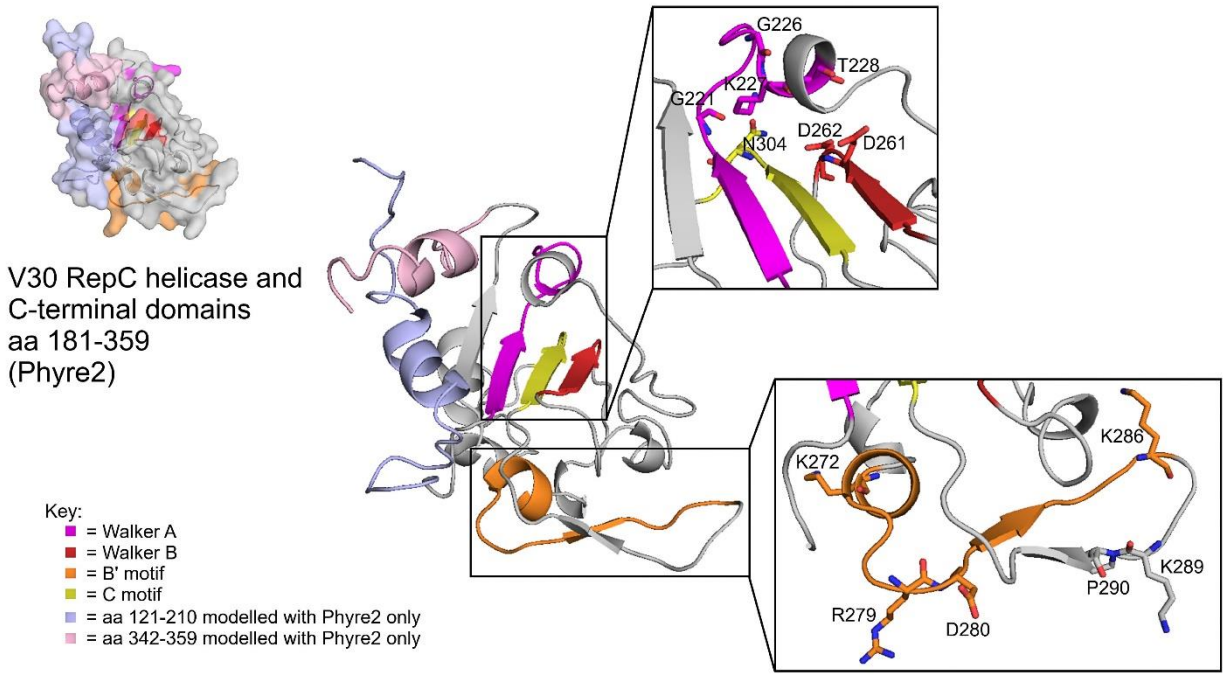


Fig. 3.7 Semi-transparent surface and cartoon representation of ToCSV V30 Rep helicase and C-terminal domain homology model monomer generated using Phyre2. Positions of functional motifs highlighted according to colour key with important Walker A, Walker B, B’ motif, B’ motif-associated (hairpin loop), and C motif residues depicted using liquorice representation. SWISS-MODEL structure quality assessment outputs shown in S3 Fig. 3.8.

Both structures were visualised in 3D space using PyMOL. The RepC-P structure (Fig. 3.7) was shown to contain six β-strands and eight helices. Four of the β-strands (β1,

β2, β5, β6) were arranged as a parallel β-sheet. The two remaining β-strands (β3 and β4) formed a distinct β-hairpin with the long loop structure pointing away from the β-sheet.

Three helices (h1, h2, h8) were situated above the β-sheet, whilst the remaining five helices were predominantly located underneath the β-sheet. The surface representation of RepC-P showed that residues belonging to WA, WB and the B' motif were positioned closely together in a putative catalytic site. WA was mapped to β1 and the N-terminal end of h2, whilst four WB residues were mapped to β2. The first five residues of the B' motif were mapped to h5 and the six C-terminal residues of the B' motif were mapped to β3 and part of the β3-assocated associated loop structure. The first five residues of the C motif formed β5. Important catalytic residues were shown using liquorice representation and labelled accordingly. Key catalytic residues present in the B' motif and associated β-hairpin were also depicted using liquorice representation (Fig. 3.7).

The homohexameric ring structure of RepC-SM contained a central channel. The six monomeric units were arranged so that the β-hairpin loop of each monomer was positioned towards the central channel (Fig. 3.8).

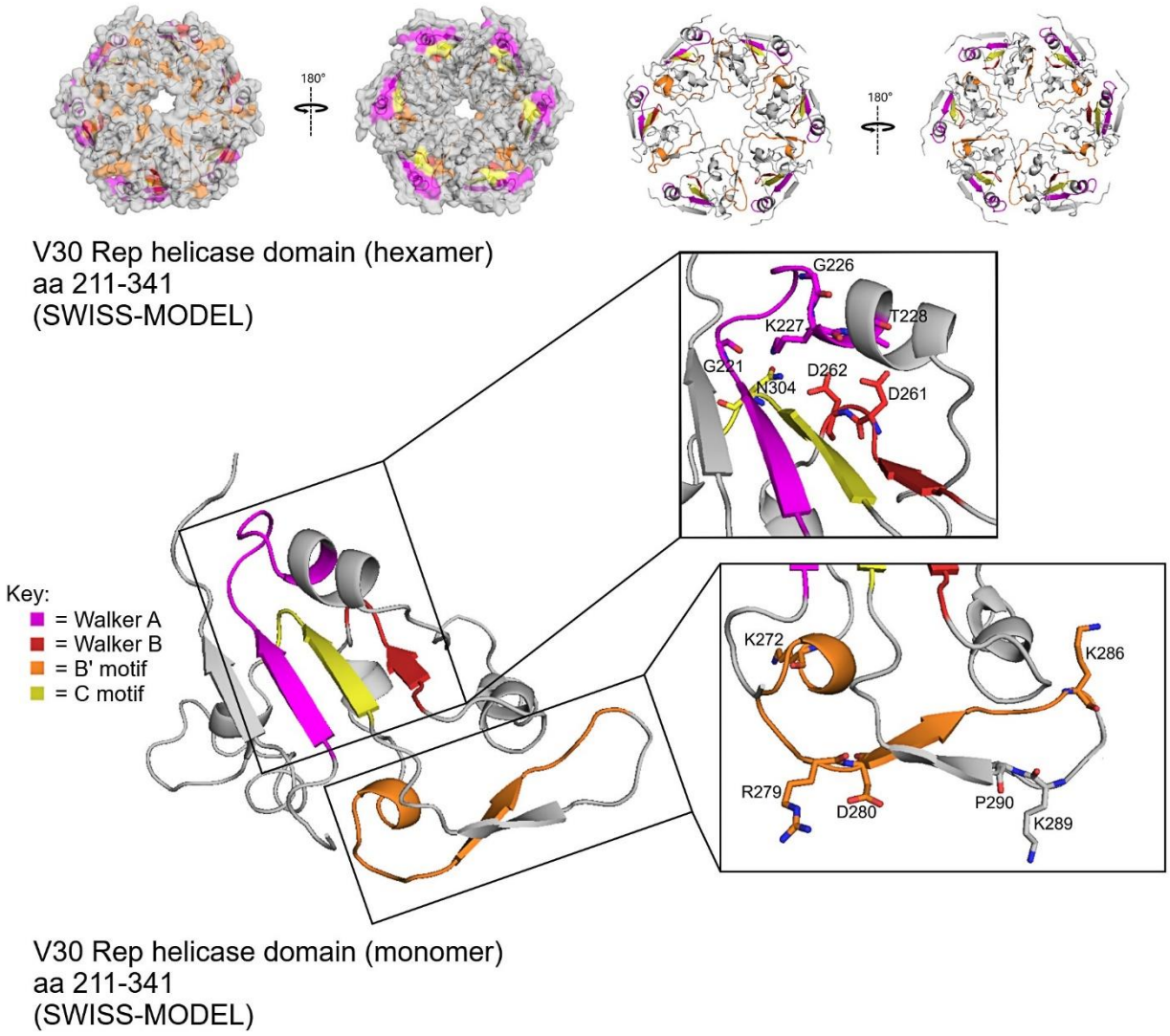


Fig. 3.8 Semi-transparent surface and cartoon representation of ToCSV V30 Rep partial helicase and C-terminal domain homology model hexamer (top) and selected representative monomer (Chain E, bottom) generated using SWISS-MODEL. Positions of functional motifs highlighted according to colour key with important Walker A, Walker B, B' motif, B' motif-associated (hairpin loop), and C motif residues depicted using liquorice representation. SWISS-MODEL structure quality assessment outputs shown in S3 Fig. 3.9.

The monomeric units of the RepC-SM hexamer are partial domain models and as a result do not contain aa sequence corresponding to h1 and h8 of the RepC-P model. The consensus ss of the monomer is shown in S3 Fig. 3.10 and was comparable to that of RepC-P, except that the consensus RepC-SM structure was predicted to contain an additional helix between β5 and β6. The arrangement of the functional motifs was very similar to that seen in RepC-P with some differences in the relative position of the key catalytic residues shown using liquorice representation (Fig. 3.8). Quality assessment of the two structures using SWISS-MODEL was done (S3 Fig. 3.8

and S3 Fig. 3.9) and the outputs showed that although RepC-SM scored slightly better than RepC-P, the overall quality of both structures was low. The Ramachandran favoured/outlier percentages were 82.2/4.9 and 80.2/9.6 for RepC-SM and RepC-P respectively. The QMEAN Z-scores obtained were more favourable for RepC-SM than RepC-P. The QMEANDisCo Global scores obtained were 0.39 ± 0.05 for RepC-SM and 0.37 ± 0.06 for RepC-P. Despite the low global scores, local quality estimates revealed that the highest scoring residues included those present in the conserved functional motifs WA, WB, and C motif; as well as residues forming β6 which is part of the C-terminal domain.

Sequence comparison of the four functional helicase domain motifs of V30 Rep with those present in the Rep protein sequences of other DNA viruses (Fig. 3.9A) revealed conservation of several important residues. WA contained consensus glycine residues at motif positions six and eleven, lysine at position twelve, and threonine at position 13. WB was shown to contain three consensus hydrophobic residues and two negatively charged residues at positions five and six. These two residues were both aspartic acid in the Rep proteins of V30, TYLCV, WDV, PCV2, and BPV (genus *Deltapapillomavirus*). The B' motif was found to be variable with only the last C-terminal lysine being conserved at 100 % consensus. The first residue of this motif was positively charged in all seven virus Rep proteins. The C motif contained three consensus hydrophobic residues and a conserved asparagine at motif residue position six. Interestingly, all three geminivirus Rep proteins lacked the arginine finger motif present in the other animal-infecting virus Rep proteins. The arginine finger is typically situated in the stretch of residues located between sequences forming β5 and β6 of the parallel β-sheet. It is worth noting that an arginine residue was identified in this stretch of residues in the Rep proteins of TbLCZV, tomato leaf curl Mali virus (ToLCMLV), tomato leaf curl Madagascar virus (ToLCMGV/Men), and ToLCUV (Fig. 3.5). The vast majority of begomovirus Rep proteins, including ToCSV Rep, contained four highly conserved positively charged lysine residues in this region (Fig. 3.5).

The PyMOL-assigned ss of RepC-P as well as the consensus monomer ss for RepC-SM and partial solved structures of other DNA virus helicase domains (S3 Fig. 3.10) were used to construct protein topology diagrams (Fig. 3.9B). The region defined using these diagrams was limited to the sequence spanning the parallel β-sheet structure. This was done to specifically compare this conserved "core" region containing

important functional motifs. The topology diagrams revealed that all structures contained a parallel β-sheet surrounded by three to five helices of variable lengths (Fig. 3.9B). The β-sheets of the solved structures contained an additional β-strand situated adjacent to the strand formed by residues corresponding to the WB motif (β2 in homology models). Interestingly, the structure of PCV2 Rep (PDB ID: 7LAR) contained a further additional β-strand downstream of the β-strand corresponding to β6 in the homology models. This additional strand was orientated antiparallel relative to the remaining strands of the of the conserved core β-sheet resulting in a mixed β-sheet. The relative positions of WA, WB and C motif were highly analogous although the β-strand lengths were variable. All structures contained the B' motif-associated β-hairpin loop structure. The N-terminal region of the B' motif forming a short helix in the homology models was found to contribute to the C-terminal end of a longer helical element in the solved structures. The B' motif catalytic residue was consistently positioned at the turn region of the β-hairpin loop in all structures. The AAV2Rep structure (PDB ID: 7JSI) contained an additional β-hairpin located downstream of the β-strand containing C motif residues. As in the case of the B' motif β-hairpin loop, the loop of the second β-hairpin was also found to be directed towards the central channel in the solved hexameric structure of AAV2 Rep. The arginine finger motif (missing in ToCSV Rep) was present between the C motif β-strand and the C-terminal parallel β-strand in the solved Rep structures.

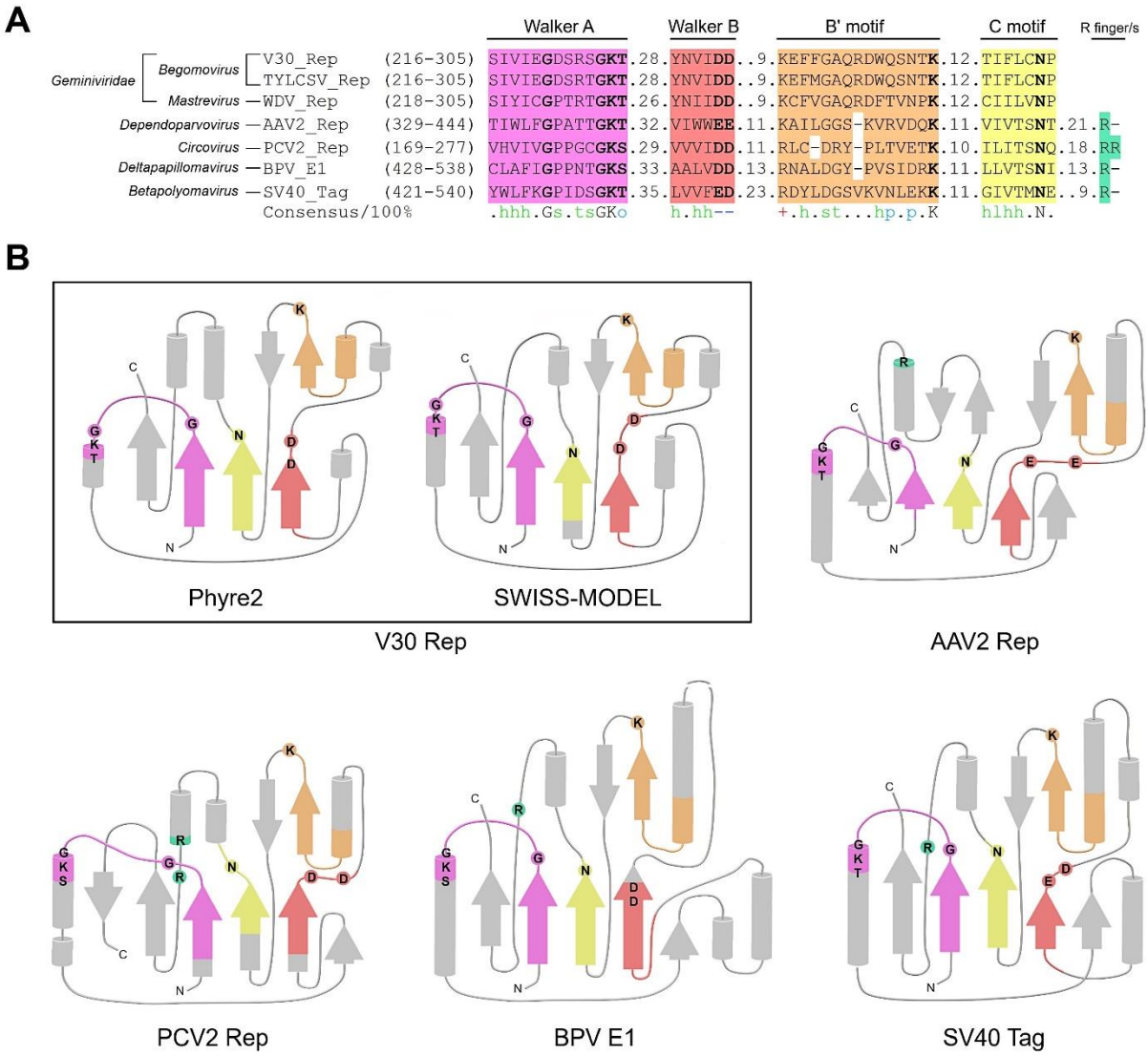


Fig. 3.9 Comparative analysis of selected DNA virus SF3 helicase C-terminal domains. **A**, MUSCLE aa sequence alignment of helicase domain functional motifs of ToCSV V30 Rep, TYLCSV Rep (NP_620741), WDV Rep (AHZ63100), AAV2 Rep (QDH44425), PCV2 Rep (AAQ94098), BPV E1 (NP_056739), and SV40 Tag (NP_043127). Motifs highlighted and annotated, with important residues in bold. Consensus residues capitalised, consensus alcohol (o), aliphatic (l), hydrophobic (h), negatively charged (-), polar (p), positively charged (+), small (s), and turnlike (t) residues indicated with coloured lowercase letters. Hyphens indicate alignment-generated gaps, ellipses in alignment indicate spacer sequence with aa number shown, protein sequence aa positions in brackets. **B**, Structural topology of V30 Rep partial helicase and C-terminal domain homology models generated using Phyre2 (monomer) and SWISS-MODEL (hexamer); and solved hexameric structures of AAV2 Rep (PDB ID: 7JSI), PCV2 Rep (PDB ID: 7LAR), BPV E1 (PDB ID: 2GXA), and SV40 Tag (PDB ID 1SVM). PyMOL-assigned "sheet" shown as arrows, "helix" shown as cylinders, "loop" shown as thin tubes. Position of Walker A motif in magenta, Walker B motif in scarlet, B' motif in orange, C motif in yellow, arginine finger motif in aquamarine. Positions of important motif residues indicated using bold letters. For all hexameric structures, consensus monomer ss is represented. Details of PyMOL-assigned ss shown in S3 Fig. 3.10.

3.3.2 The disordered N- and C-termini of V30 Rep warrant increased scrutiny for their role in the functionality of Rep.

Upon preliminary alignment of African begomovirus Rep protein sequences obtained following standard code using ORF finder, it was discovered that several identified Rep proteins were predicted to encode additional N-terminal segments. These N-terminal extensions were situated upstream of a methionine conserved amongst all African begomovirus Rep proteins analysed. These N-terminal extensions were only identified in monopartite begomoviruses. These unique proteins are herein referred to as “Rep-N”. All three ToCSV sequences were predicted to encode Rep-N proteins each containing an additional 12 aa upstream of the conserved methionine. A sequence alignment of several identified Rep-N proteins showed that the N-terminal extension of TbLCZV Rep-N was identical to that of ToCSV (Fig. 3.10A).

To investigate whether the presence of this putative segment altered the predicted ss/disorder of V30 Rep, the sequences of Rep-N and canonical Rep proteins were analysed further. It was determined that ss predicted at medium to high confidence levels (≥5) for the two proteins was nearly identical (S3 Fig. 3.11 and S3 Fig. 3.12). Predicted disorder, however, was markedly different (Fig. 3.10 B). The canonical V30 Rep protein was predicted to contain three disordered regions with residues having DISOPRED scores above the disorder threshold of 0.5. These included four residues in the IRD of the N-terminal domain (alignment position 13-16), eight in the oligomerisation domain (alignment position 134-141), and 16 in the C-terminal domain (alignment position 356-371). All disordered residues were predicted to have protein binding potential except for asparagine at alignment position 141. The predicted N-terminal region disorder of Rep-N was extended to include the entire putative fragment (alignment positions 1-12). The predicted disorder of the oligomerisation domain region seen in canonical Rep was lost in Rep-N as the DISOPRED scores of the corresponding residues were below the 0.5 threshold value. The length of the disordered region present in the C-terminal domain of Rep-N was reduced by one residue in comparison to canonical Rep. To probe for the presence of a cs strand-derived RNA transcript corresponding to the 5’ C1 N-terminal extension region, RT-

PCR was done using C1-specific primers. The expected product was positively identified in plants infected with V30, but not mock-inoculated plants (Fig. 3.10C).

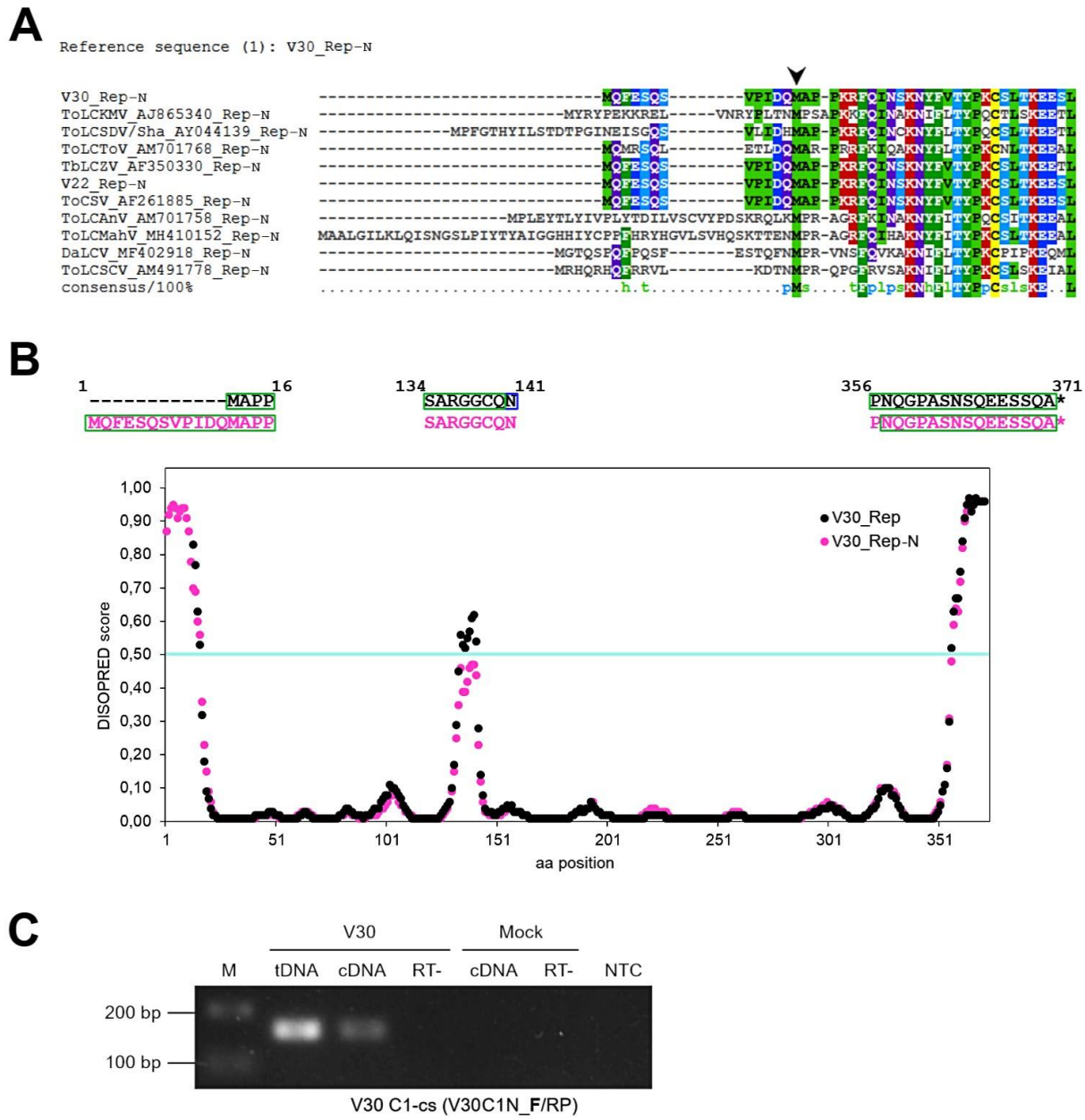


Fig. 3.10 Analysis of putative begomovirus Rep N-terminal region extension. **A**, Multiple Rep-N aa sequence alignment (MUSCLE) for ToCSV V30, V22, exemplar isolate (AF261885), and selected African monopartite begomoviruses having C1 ORF encoding putative Rep N-terminal segment predicted by ORF finder. Residues identical to reference sequence (V30 Rep-N) highlighted. Canonical Rep aa 1 (methionine) indicated with black arrow. Consensus aliphatic (l), hydrophobic (h), polar (p), small (s) and turnlike (t) residues indicated with coloured lowercase letters. **B**, Comparative DISOPRED analysis of canonical V30 Rep and Rep containing putative N-terminal extension (Rep-N, pink sequence). Disordered residues boxed in blue, disordered residues with protein-binding potential boxed in green. Disorder threshold value (0.5) demarcated with turquoise line. Residue positions indicated. **C**,

Agarose gel image of V30 complementary sense (cs) putative C1 ORF 5' extension region-derived RT-PCR products (encoding putative Rep N-terminal extension). M, DNA ladder marker; tDNA, PCR of total DNA extracted from plants inoculated with V30 (positive control); cDNA, RT-PCR of total RNA extracted from plants inoculated with V30 or mock; RT-, negative reverse transcriptase enzyme control; NTC, PCR no template control. Primers used for PCR in brackets with primer used for cDNA synthesis highlighted in bold. Mock is RT-PCR of total RNA extracted from plants inoculated with *Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300. Rep sequences obtained with ORF finder using the DNA-A sequence as the input. DNA-A sequences for all isolates used in S3 Table 3.2.

To explore the role that the C-terminal domain of V30 Rep may play in viral infectivity, two C1 deletion mutants [V30ΔC1-d(14) and V30ΔC1-d(22)] were generated. The predicted ss and disorder of the mutated C1 proteins was compared with the wild type V30 Rep. At medium to high confidence levels (≥5) the predicted ss of the three proteins was identical (S3. Fig. 3.11, S3 Fig. 3.13, S3 Fig. 3.14). Truncation of the Rep protein by 22 aa resulted in the removal of a predicted helix at Rep aa 339-341, although this helix was predicted with low confidence (<5). Predicted disorder was substantially altered by truncation of the C-terminus (Fig. 3.11).

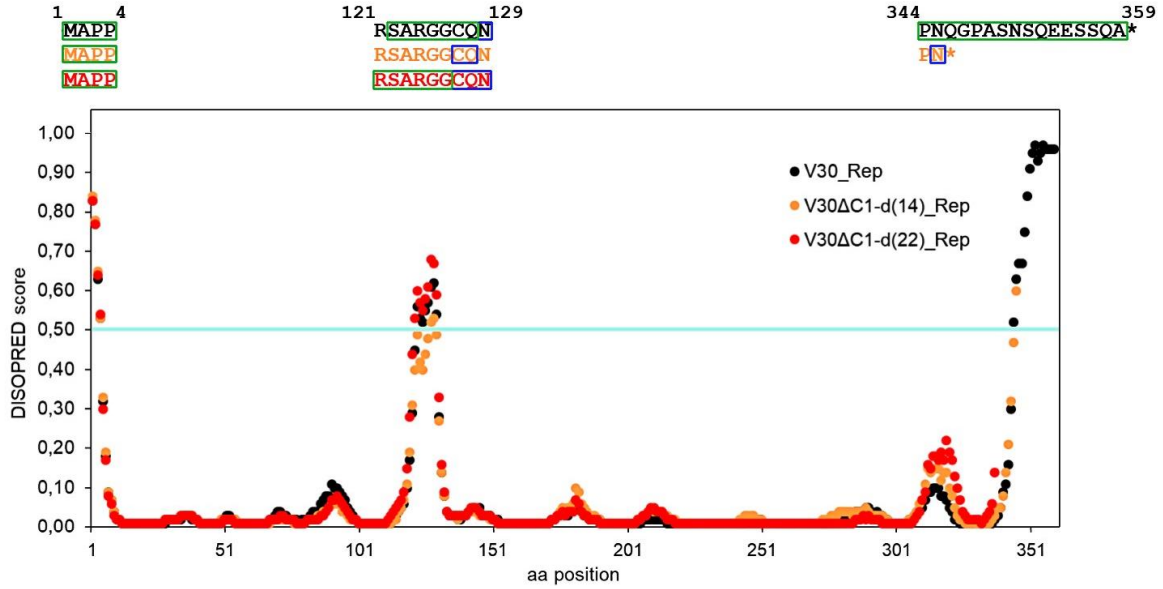


Fig. 3.11 Comparative DISOPRED analysis of canonical ToCSV V30 Rep, V30ΔC1-d(14) Rep (orange sequence), and V30ΔC1-d(22) Rep (red sequence). Disordered residues boxed in blue, disordered residues with protein-binding potential boxed in green. Disorder threshold value (0.5) demarcated with turquoise line. Residue positions indicated.

Truncation of Rep by 14 residues resulted in a near total loss of the C-terminal disordered region having protein binding potential. Only the C-terminal asparagine at

position 345 was predicted to be disordered. Deletion of 22 C-terminal Rep residues resulted in a complete loss of the C-terminal region of disorder. The predicted disorder present in the oligomerisation domain of V30 Rep was significantly reduced in the Rep of V30ΔC1-d(14) and its predicted protein binding potential was lost. Conversely, the Rep of V30ΔC1-d(22) was predicted to retain this region of disorder and was expanded to include arginine at position 121. The protein binding potential of the disordered residues cystine (position 127) and glutamine (position 128) in the Rep of V30ΔC1-d(22) was lost. The predicted disorder in the N-terminal domain of V30 Rep was not altered by the truncation of the C-terminus.

Infectivity trials were conducted in *N. benthamiana* using V30ΔC1-d(14) and V30ΔC1-d(22) to determine whether truncation of the Rep C-terminal domain interfered with disease development. Plants inoculated separately with the two mutants did not develop any disease symptoms up to 36 dpi and were indistinguishable from mock-inoculated plants (Fig. 3.12A). At ~40 dpi however, plants infected with V30ΔC1-d(14) began to show slight SV symptoms in the emerging leaves. By 50 dpi the SV symptoms increased in severity and leaf rugosity was evident. Impaired flower development was also observed, unlike in the mock and V30ΔC1-d(22)-inoculated plants which continued to grow and flower normally (Fig. 3.12B).

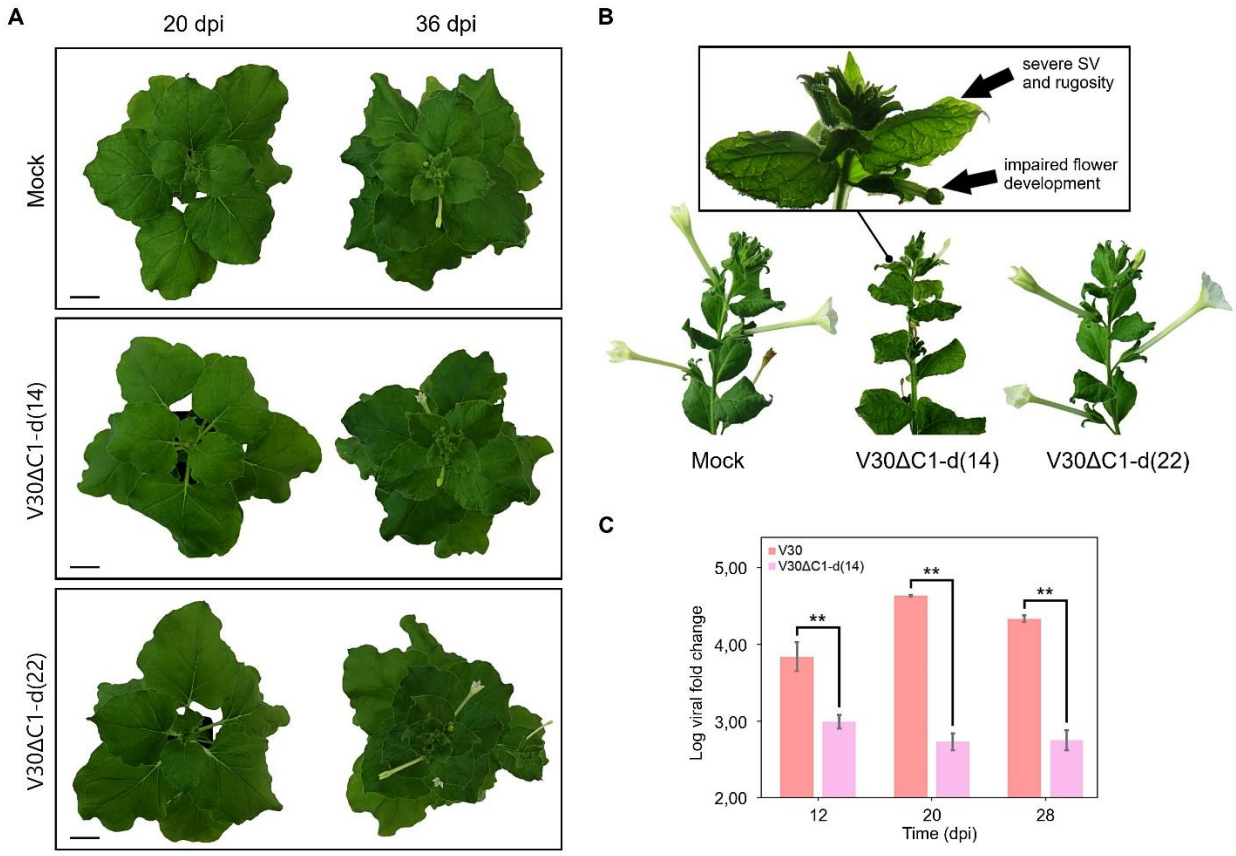


Fig. 3.12 Infection of *Nicotiana benthamiana* with ToCSV V30 C1 mutant infectious clones in single infection. **A**, Symptoms in plants inoculated with mock (*Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300), V30ΔC1-d(14), and V30ΔC1-d(22) at 20 dpi (left) and 36 dpi (right). Bar = 2 cm. **B**, Symptoms in plants inoculated with mock, V30ΔC1-d(14), and V30ΔC1-d(22) at ~50 dpi. Virus-like symptoms observed only in V30ΔC1-d(14)-infected plants seen magnified (inset). **C**, Log viral fold change at 12, 20, and 28 dpi in V30 and V30ΔC1-d(14)-infected plants. Bars represent mean ± SD. Student's t-test level of significance: **, P < 0.01.

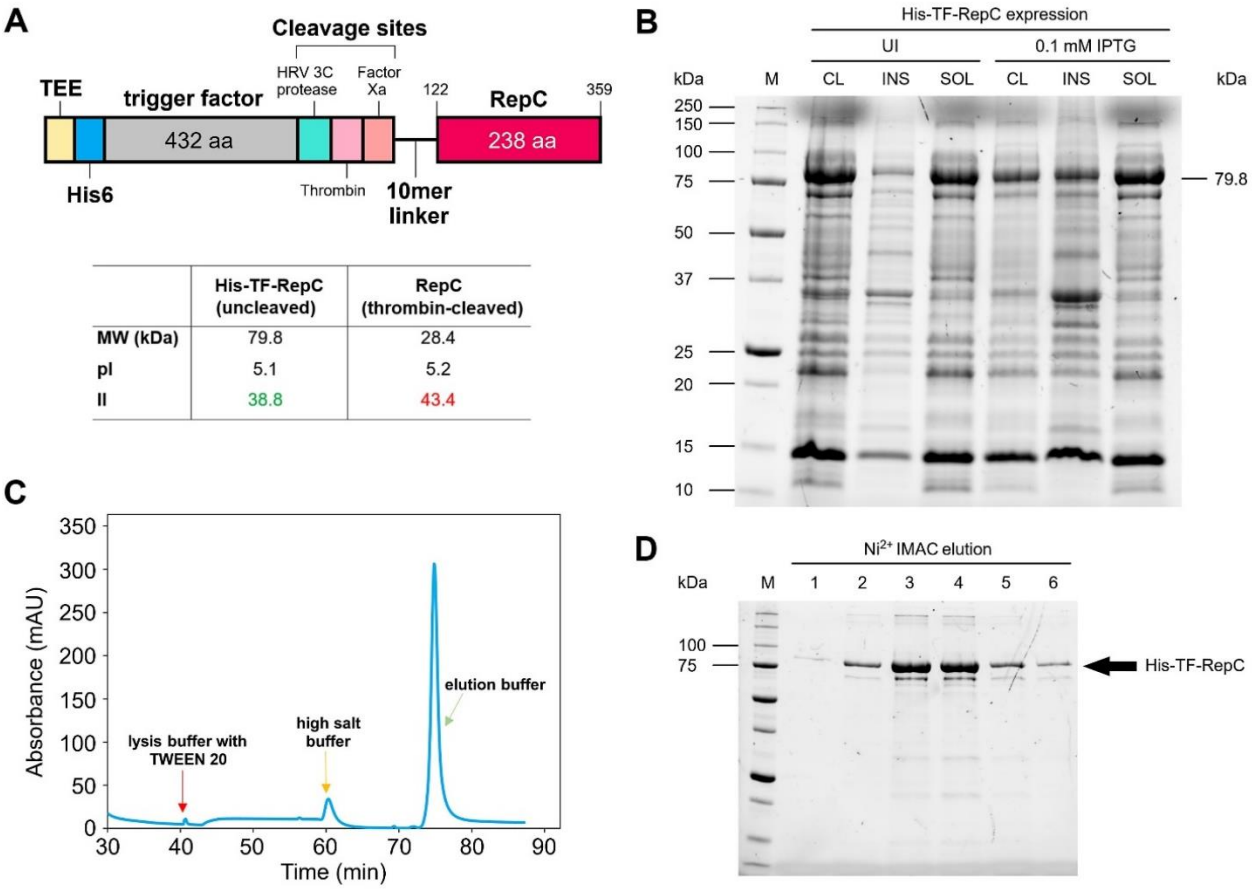
DNA was extracted from plants inoculated with V30ΔC1-d(14) at ~50 dpi and samples were found to be PCR-positive (V2 region primers were used). Sanger sequencing of the samples indicated that mutation reversion had not occurred, and that the integrity of the mutant was retained. Repeated PCR testing of V30ΔC1-d(22)-inoculated plants did not yield a positive PCR result. Despite not displaying symptoms, plants inoculated with V30ΔC1-d(14) were found to contain measurable levels of viral DNA. Analysis of viral load at 12, 20 and 28 dpi showed that V30ΔC1-d(14)-infected plants had a significantly lower viral load than plants infected with V30 at all three timepoints ($P < 0.01$) (Fig. 3.12C). In addition, the peak in viral load observed at 28 dpi in V30-infected plants was not evident in V30ΔC1-d(14)-infected plants. The viral load in mutant-infected plants did not significantly differ between the three timepoints.

3.3.3 Expression and purification of recombinantly expressed V30 RepC proteins – a complex puzzle with many pieces.

Although published structures of the Rep helicase domain of several animal-infecting DNA viruses exist in the PDB, the structure of the geminivirus Rep helicase domain has not yet been solved. To date, only geminivirus Rep N-terminal domain structures have been published. The first step on the path to solve a particular molecular structure is to obtain significant amounts of purified target protein. In the current study, the DNA sequence encoding the helicase (aa 122-330) and C-terminal (aa 330-359) domains of V30 Rep (termed “RepC”) were cloned into three *E. coli* expression vectors and expression trials were conducted. The pCold TF expression vector allowed for expression of RepC with a trigger factor (TF) chaperone tag fused to the N-terminal end of the protein to assist in the folding of translated RepC. A map of the recombinantly expressed protein is shown in Fig. 3.13A. The recombinant protein contained a hexahistidine (His6) tag upstream of the TF tag to facilitate purification using IMAC. Three protease cleavage sites were present between the TF tag and the RepC protein to enable tag separation following purification. The theoretical pI values of His-TF-RepC and thrombin-cleaved RepC were nearly identical whilst the predicted instability index (II) values for the two proteins differed (Fig. 3.14A). The recombinant fusion was predicted to be stable in solution (II: 38.8) and the cleaved RepC protein was predicted to be unstable because the II value (43.4) exceeded the instability threshold value of 40.0.

The recombinant protein (His-TF-RepC) was successfully overexpressed predominantly in the soluble fraction using BL21(DE3)pLysS. This was confirmed using SDS-PAGE by the presence of a distinct band at ~80 kDa corresponding to the predicted molecular mass of the His-TF-RepC protein (79.8 kDa) (Fig. 3.13B). Basal level expression was also observed in the absence of IPTG. The wet cell pellet yield obtained was ~1 g per 100 mL culture.

The soluble fraction was subjected to purification using IMAC with immobilised Ni^{2+} resin. Fractions corresponding to the absorbance (Abs) peak obtained after application of the elution buffer (Fig. 3.13C) were shown to contain the target protein using SDS-PAGE (Fig. 3.13D). The apparent purity of the elution fractions was improved by subjecting the column to detergent and high salt washes prior to target protein elution.



	His-TF-RepC (uncleaved)	RepC (thrombin-cleaved)
MW (kDa)	79.8	28.4
pI	5.1	5.2
II	38.8	43.4

Fig. 3.13 Expression and purification of His-TF-RepC protein. **A**, Map of recombinant His-TF-RepC as expressed using pCold TF-RepC. Translation enhancing element (TEE) in yellow, hexahistidine (His6) tag in blue, trigger factor (TF) chaperone tag in grey. Three cleavage sites indicated: HRV 3C protease in aquamarine, thrombin in pink, factor Xa in orange. A 10mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved His-TF-RepC and thrombin-cleaved RepC indicated below map. II in green indicates stable, II in red indicates unstable. **B**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced cell lysate (CL), insoluble (INS), and soluble (SOL) fractions extracted from *E. coli* BL21(DE3) pLysS liquid culture transformed with pCold TF-RepC after overexpression at 15 °C for 24 h. Target product band (79.8 kDa) indicated. **C**, Absorbance profile of one-step Ni^{2+} IMAC ÄKTA purification of His-TF-RepC showing peaks corresponding to TWEEN 20 wash, high salt wash, and target elution with buffer containing 500 mM imidazole. **D**, SDS-PAGE analysis of six sequential elution fractions obtained from one-step Ni^{2+} IMAC ÄKTA purification of His-TF-RepC (indicated using arrow) from soluble fraction extracted from *E. coli* BL21(DE3) pLysS liquid culture transformed with pCold TF-RepC after overexpression at 15 °C for 24 h. M, protein molecular mass marker. Details of nt and predicted aa sequences of recombinantly expressed His-RF-RepC in S3 Fig. 3.1.

A proteolytic cleavage experiment to separate the His-TF tag from RepC was conducted using thrombin (Fig. 3.14A). SDS-PAGE confirmed successful tag cleavage by the gradual decrease in the intensity of a band at ~80 kDa (His-TF-RepC) with increasing cleavage time. An associated increase in the intensity of bands corresponding to cleaved His-TF tag (51.4 kDa) and cleaved RepC (28.4 kDa) was recorded with near total cleavage of the His-TF tag being obtained after ~6 h. The presence of additional non-target cleavage products at low molecular mass (~15 kDa) was also shown to increase over time although their sequence identity was not determined. SDS-PAGE of flowthrough and elution fractions obtained from a second IMAC purification of cleaved protein (Fig. 3.14B) showed a band at ~30 kDa (cleaved RepC) as well as a non-target low molecular mass cleavage product. The elution fraction contained a faint band at ~80 kDa (uncleaved His-TF-RepC) and a distinct band at ~50 kDa (cleaved His-TF tag). The presence of a band at ~30 kDa (cleaved RepC) was also identified with an apparent band intensity higher than that seen in the flow through fraction. Low molecular mass non-target cleavage products were also observed in the elution fraction.

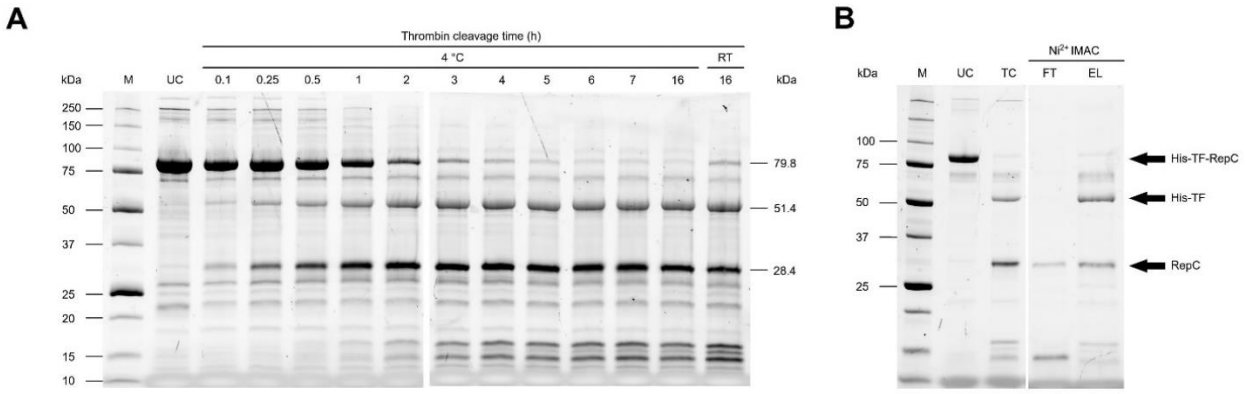
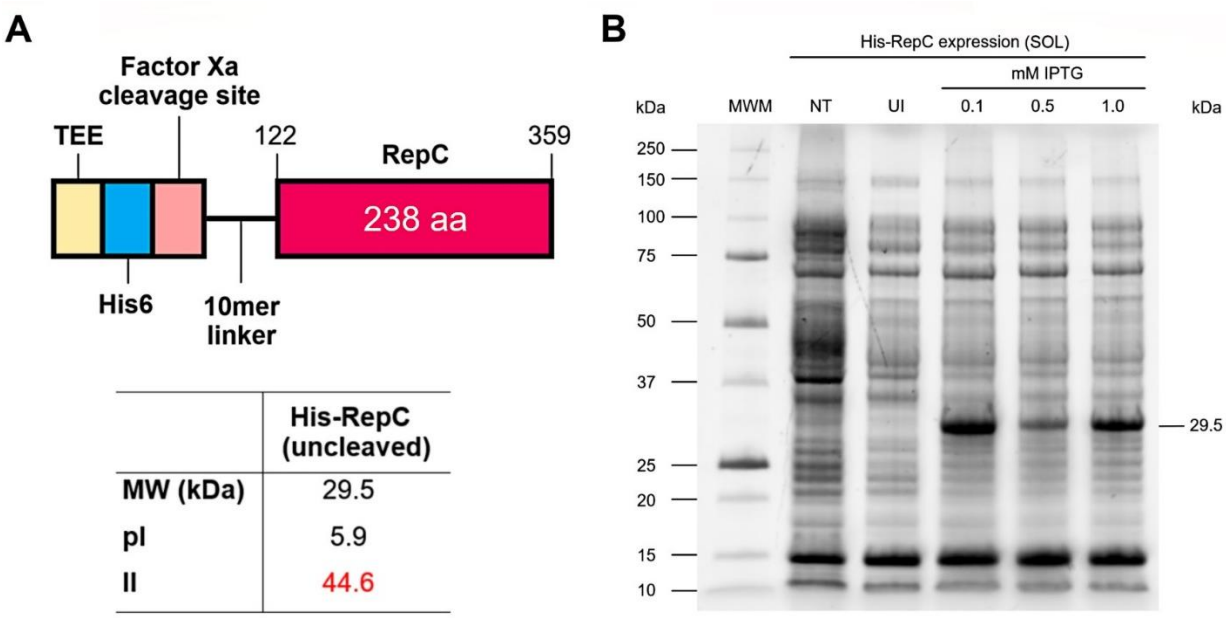


Fig. 3.14 Thrombin cleavage of His-TF-RepC and subsequent purification of RepC protein. **A**, SDS-PAGE analysis of concentrated Ni^{2+} IMAC ÄKTA-purified His-TF-RepC subjected to cleavage with 1 U thrombin at 4 °C for up to 16 h. Cleavage reaction samples were analysed at indicated time intervals. Uncleaved (UC) His-TF-RepC (79.8 kDa), cleaved His-TF (51.4 kDa) tag and RepC (28.4 kDa) bands indicated. **B**, SDS-PAGE analysis of UC and thrombin-cleaved (TC) His-TF-RepC; as well as flow-through (FT) and elution (EL) fractions obtained from one-step Ni^{2+} IMAC ÄKTA purification of RepC after thrombin cleavage for 16 h at 4 °C. UC His-TF-RepC, TC His-TF, and RepC indicated using arrows. M, protein molecular mass marker.

The pCold I expression vector, which uses the same expression technology as pCold TF but lacks the TF tag, was used to recombinantly express RepC in T7 Express. This

expression strain was used since other strains transformed with this modified expression vector did not grow correctly or did not yield any recombinant protein in either the soluble or insoluble fractions. The recombinant protein expressed using this vector (His-RepC) contained a His6 tag and a factor Xa cleavage site upstream of RepC (Fig. 3.15A). The II value of His-RepC (Fig. 3.15A) classified this protein as unstable in solution. The His-RepC protein was expressed in the soluble fraction as evidenced by the presence of an SDS-PAGE band at ~30 kDa corresponding to the predicted molecular mass of His-RepC (29.5 kDa) (Fig. 3.15B). Unlike expression of RepC using pCold TF, the expression of RepC using pCold I in the soluble fraction was inconsistent. Additionally, wet cell pellet yield (~0.35g per 100 mL culture) was greatly reduced in comparison.



	His-RepC (uncleaved)
MW (kDa)	29.5
pI	5.9
II	44.6

Fig. 3.15 Expression of His-RepC protein. **A**, Map of recombinant His-RepC as expressed using pCold I-RepC. Translation enhancing element (TEE) in yellow, hexahistidine (His6) tag in blue. Factor Xa cleavage site in orange. A 10mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved His-RepC indicated below map. II in red indicates unstable. **B**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced soluble (SOL) fractions extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. Target product band (29.5 kDa) indicated. Non-transformed (NT) soluble fraction included for comparison. M, protein molecular mass marker. Details of nt and predicted aa sequences of recombinantly expressed His-RepC in S3 Fig. 3.2.

Three separate methods were used in the purification of His-RepC extracted in the soluble fraction (Fig. 3.16). IMAC using immobilised Ni^{2+} resin was conducted, and SDS-PAGE analysis revealed the presence of a band at ~30 kDa (His-RepC) predominantly in the flow through and initial wash fraction obtained. The elution fractions were not pure and appeared to contain low amounts of the target protein (Fig. 3.16A). AS precipitation was done using soluble fraction containing His-RepC to "salt out" the target protein. SDS-PAGE analysis of supernatant and pellet samples obtained at a range of AS concentrations (percentage saturation) is presented in Fig. 3.16B. It was found that a distinct band at ~30 kDa (His-RepC) was present in the pellet fraction at 35 % saturation with this sample having improved purity in comparison to that of the sample at 0 % saturation. The third purification method utilised was gel filtration chromatography with sepharose 6B agarose. SDS-PAGE of all collected elution fractions revealed the presence of a distinct band at ~30 kDa (His-RepC) in elution fractions 7-11 with significantly improved purity being evident (Fig. 3.16C).

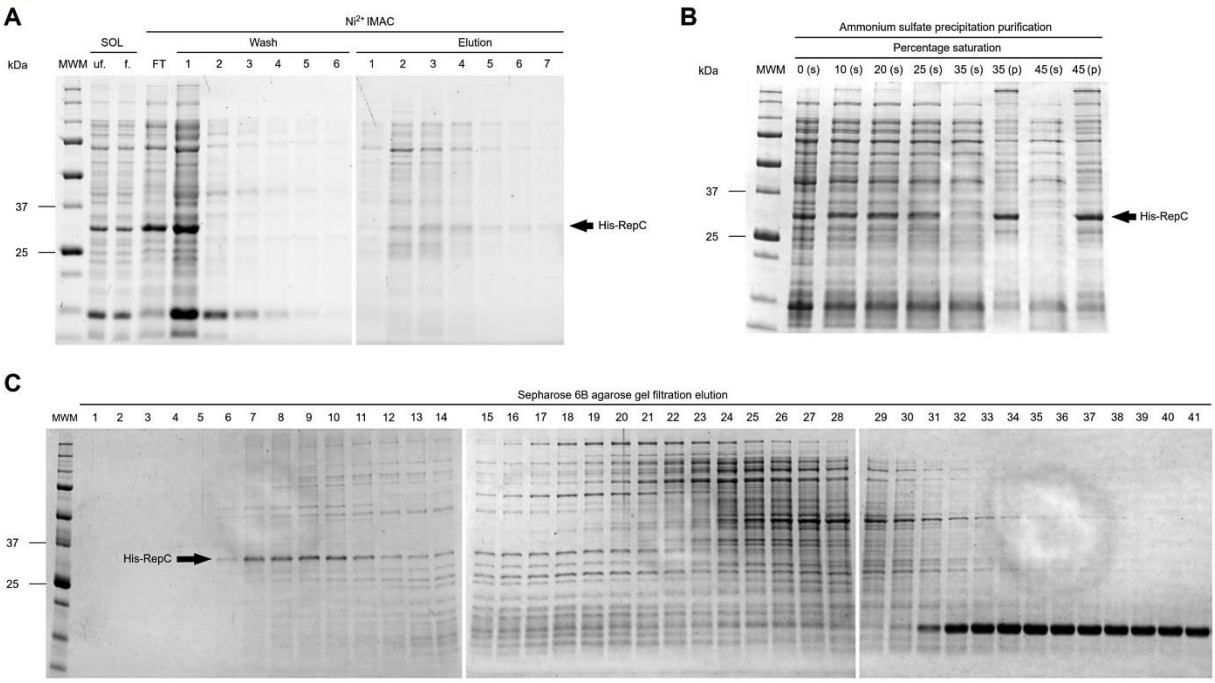
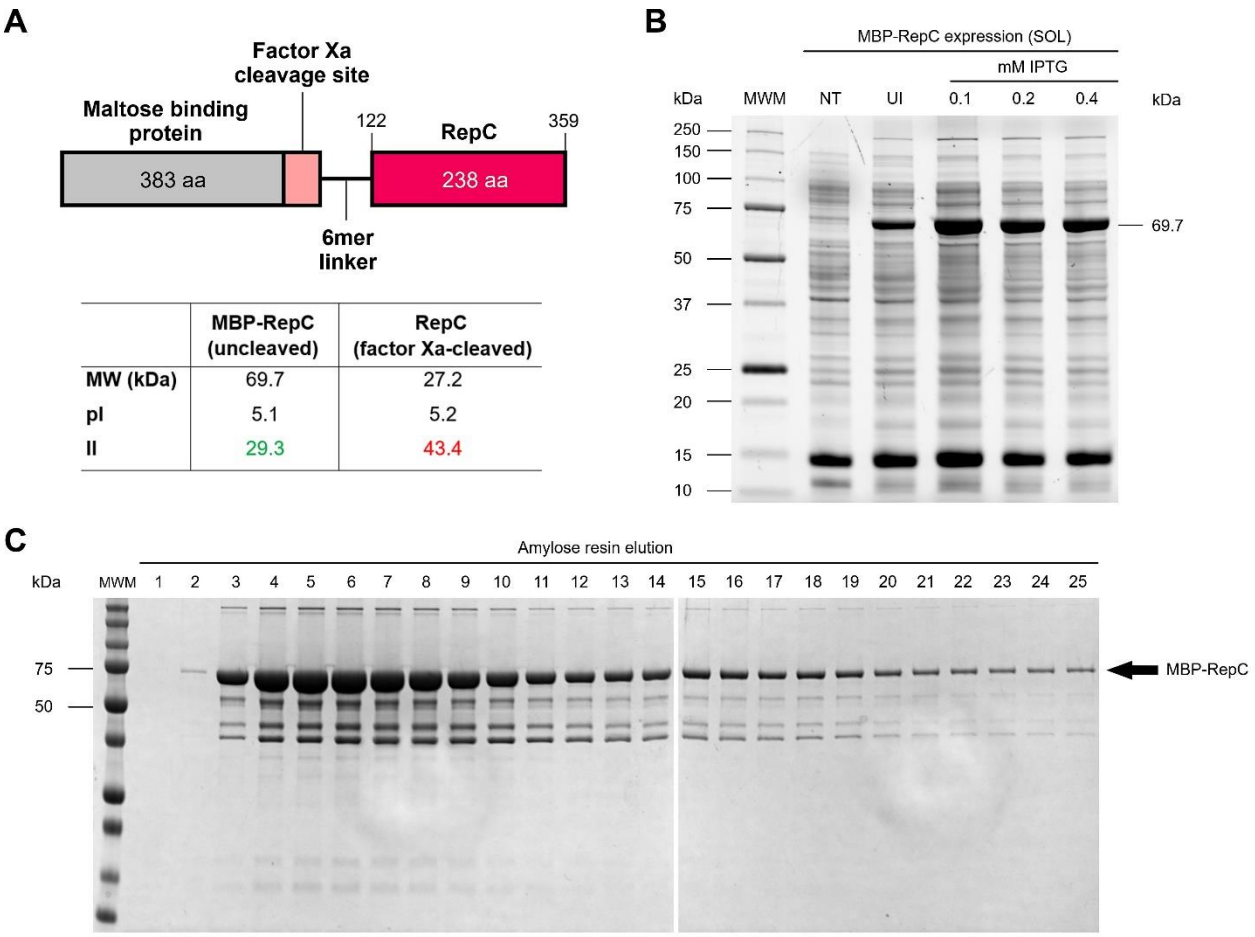


Fig. 3.16 Purification of His-RepC protein. **A**, SDS-PAGE analysis of pooled unfiltered (uf.) and filtered (f.) extracted soluble fraction; as well as flow-through (FT), six sequential wash, and six sequential elution fractions obtained from Ni^{2+} IMAC gravity flow purification of His-RepC (indicated using arrow) from filtered soluble fraction extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. **B**, SDS-PAGE analysis of supernatant (s) and

resuspended pellet (p) fractions obtained from AS purification of His-RepC (indicated using arrow) from filtered soluble fraction extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. Varying AS final concentrations (percentage saturation) used indicated. **C**, SDS-PAGE analysis of 41 sequential elution fractions obtained from sepharose 6B agarose gel filtration gravity flow purification of His-RepC (indicated using arrow) from filtered soluble fraction extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. M, protein molecular mass marker.

The pMAL-c2X expression vector allowed for expression of RepC with a maltose binding protein (MBP) tag fused to the N-terminal end of RepC to improve solubility and facilitate purification using amylose affinity chromatography. A map of the recombinantly expressed protein (MBP-RepC) is shown in Fig. 3.17A. The MBP-RepC protein contained a factor Xa cleavage site between the MBP tag and RepC to allow for tag removal after purification. The theoretical pI values of MBP-RepC and factor Xa-cleaved RepC were nearly identical whilst the predicted II values for the two proteins differed (Fig. 3.17A). The recombinant fusion was predicted to be stable in solution (II: 29.3) whilst the cleaved RepC protein was predicted to be unstable (II: 43.4). Overexpression of the recombinant MBP-RepC protein in the soluble fraction was successfully obtained using T7 Express. A distinct band at ~70 kDA corresponding to the molecular mass of MBP-RepC (69.7 kDa) was observed in the extracted soluble fraction (Fig. 3.17B). Basal level expression in the absence of IPTG was also obtained. The wet cell pellet yield was ~0.6 g per 100 mL culture.



	MBP-RepC (uncleaved)	RepC (factor Xa-cleaved)
MW (kDa)	69.7	27.2
pI	5.1	5.2
II	29.3	43.4

Fig. 3.17 Expression and purification of MBP-RepC protein. **A**, Map of recombinant MBP-RepC as expressed using pMAL-c2X-RepC. Maltose binding protein (MBP) tag in grey. Factor Xa cleavage site in orange. A 6mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved MBP-RepC and factor Xa-cleaved RepC indicated below map. II in green indicates stable, II in red indicates unstable. **B**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced soluble (SOL) fractions extracted from *E. coli* T7 Express liquid culture transformed with pMAL-c2X-RepC after overexpression at 18 °C for 16 h. Target product band (69.7 kDa) indicated. Non-transformed (NT) soluble fraction included for comparison. **C**, SDS-PAGE analysis of 25 sequential elution fractions obtained from amylose resin gravity flow purification of MBP-RepC (indicated using arrow) from filtered soluble fraction extracted from *E. coli* T7 Express liquid culture transformed with pMAL-c2X-RepC after overexpression at 18 °C for 16 h. M, protein molecular mass marker. Details of nt and predicted aa sequences of recombinantly expressed MBP-RepC in S3 Fig. 3.3.

Amylose affinity chromatography with gravity flow was used to purify MBP-RepC present in the soluble fraction. An SDS-PAGE band at ~70 kDa (MBP-RepC) was identified in the fractions obtained after elution with buffer containing 10 mM maltose

(Fig. 3.17C), although other non-target proteins were also present at lower relative amounts.

Despite the presence of these contaminants, a tag cleavage trial using factor Xa was conducted (Fig. 3.18A). SDS-PAGE analysis of cleavage samples collected at specific time points showed that tag cleavage was successful as evidenced by the decreasing intensity of the band at ~70 kDa (uncleaved MBP-RepC) over time. Concurrently, the intensity of bands corresponding to cleaved MBP tag (42.5 kDa) and cleaved RepC (27.2 kDa) increased over time. Near total tag cleavage was observed after 24 h.

Cleaved protein was subjected to ion-exchange chromatography using a DEAE FF anion exchanger column with a continuous NaCl gradient elution. During elution, two absorbance (A280) peaks were recorded (Fig. 3.18B). SDS-PAGE analysis (Fig. 3.18C) of elution fractions corresponding to the first absorbance peak obtained at ~120 mM NaCl (elutions 6-16) revealed the presence of a band at ~40 kDa (cleaved MBP tag). Elution fractions associated with the lower second peak obtained at ~240 mM NaCl (elutions 17-25) contained bands at ~70 kDa (uncleaved MBP-RepC) and ~28 kDa (cleaved RepC). An additional faint band at ~58 kDa was also observed although the identity of this band was not determined.

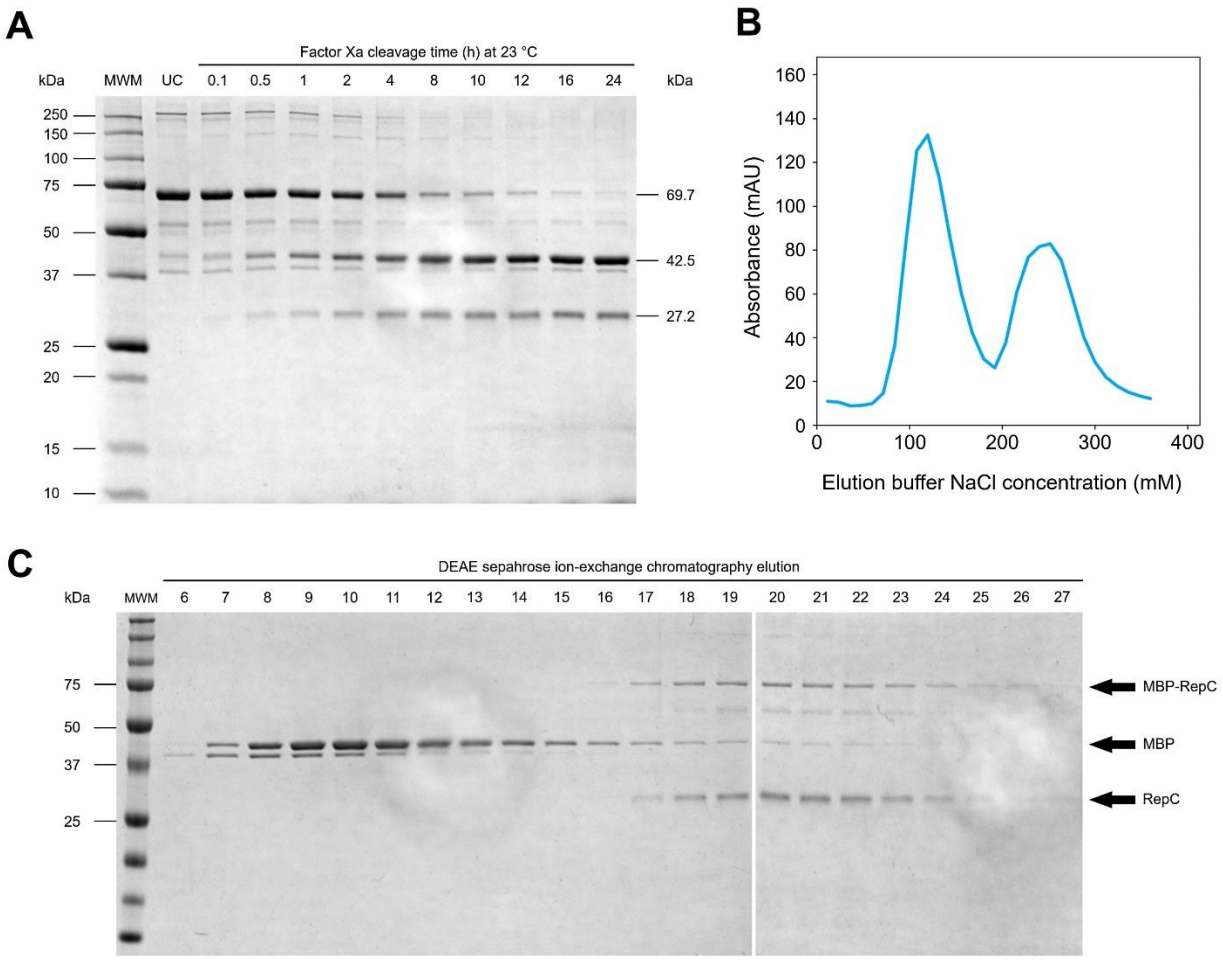


Fig. 3.18 Factor Xa cleavage of MBP-RepC and subsequent purification of RepC protein. **A**, SDS-PAGE analysis of amylose resin gravity flow-purified MBP-RepC subjected to cleavage with 2 % w/w factor Xa at 23 °C for up to 24 h. Cleavage reaction samples were analysed at indicated time intervals. Uncleaved (UC) MBP-RepC (69.7 kDa), cleaved MBP (42.5 kDa) tag, and RepC (27.2 kDa) bands indicated. **B**, Absorbance profile of ion-exchange chromatography purification of factor Xa-cleaved RepC showing peaks corresponding to continuous gradient elution with increasing NaCl concentration. **C**, SDS-PAGE analysis of 22 elution fractions obtained from ion-exchange chromatography purification of RepC after factor Xa cleavage at 23 °C for 16 h. Uncleaved MBP-RepC, cleaved MBP, and RepC indicated using arrows. M, protein molecular mass marker.

3.3.4 Miniaturised expression of truncated RepC proteins – a practical approach to improve yield and solubility.

Recombinant protein expression often requires optimisation especially in the case of proteins such as RepC which are not easily expressed/purified in large quantities and are predicted to be unstable in solution. One useful method of enhancing recombinant expression is to truncate the target protein to remove disordered regions that could

contribute to overall instability in solution. The C-terminal domain containing disordered residues was selected as a target for truncation of recombinantly expressed RepC. The mutated expression vectors pCold I-RepC_14 and pMAL-c2X-RepC_14 were generated to express a mutant protein lacking the 14 C-terminal residues (protein termed “RepC_14”). A miniaturised expression system using 1 mL expression culture volumes was designed to allow rapid and repeated comparison of the expression of wild type and mutant recombinant proteins.

The miniaturised expression trials were conducted using T7 Express because the BL21(DE3) and BL21(DE3)pLysS strains transformed with pCold I-RepC_14 did not grow correctly. SDS-PAGE analysis showed that both His-RepC and His-RepC_14 were expressed in the insoluble fraction (Fig. 3.19A, B), although the observed intensity of the SDS-PAGE band corresponding to the predicted molecular mass of His-RepC_14 (28.1 kDa) was significantly greater than that of His-RepC.

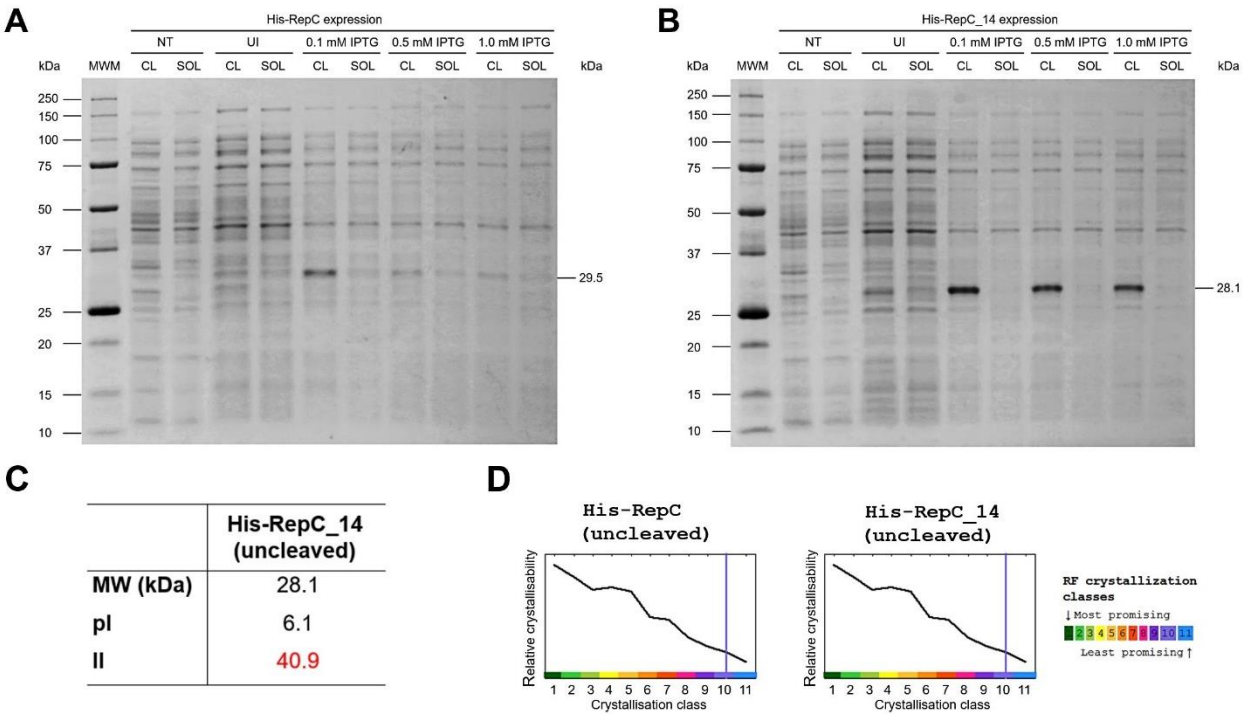
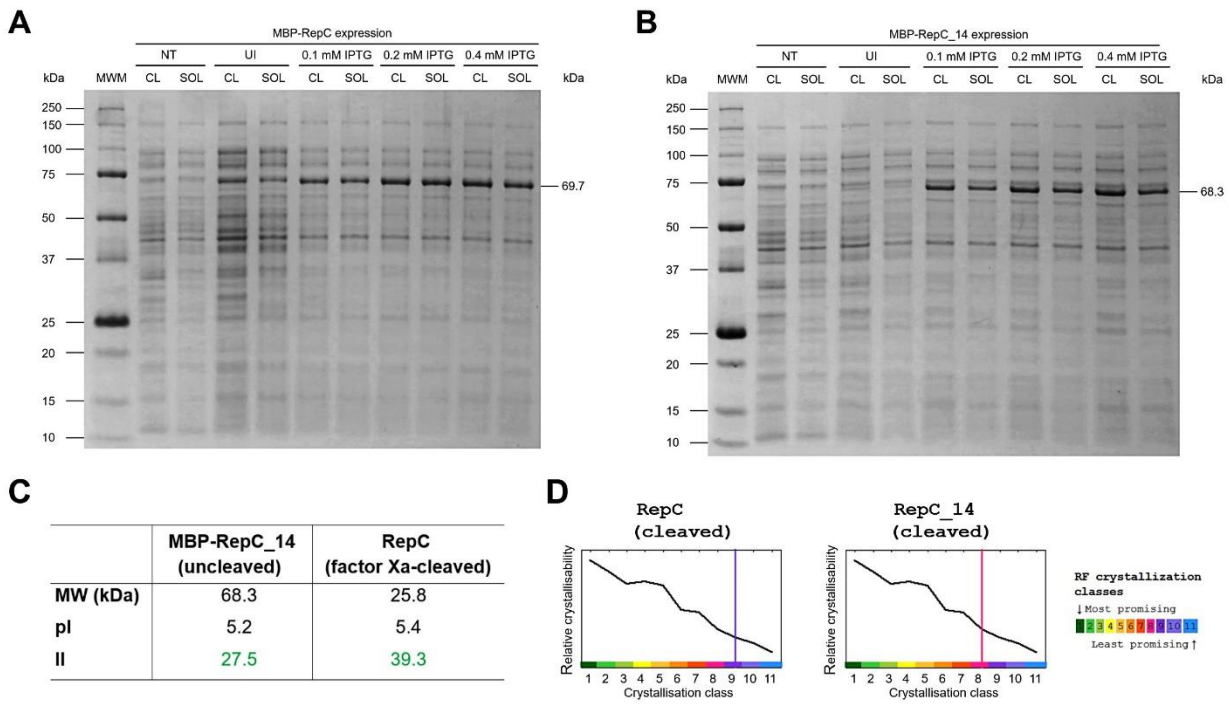


Fig. 3.19 Comparative miniaturised expression trials of His-RepC and His-RepC_14 proteins. **A**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced cell lysate (CL) and soluble (SOL) fractions extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC after overexpression at 15 °C for 24 h. Target product band (29.5 kDa) indicated. **B**, SDS-PAGE analysis of UI and IPTG-induced cell lysate CL and SOL fractions extracted from *E. coli* T7 Express liquid culture transformed with pCold I-RepC_14 (encoding RepC truncated by 14 C-terminal residues), after overexpression at 15 °C for 24 h. Target product band (28.1 kDa) indicated. Non-transformed (NT) CL and SOL fractions

included for comparison. M, protein molecular mass marker. **C**, Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved His-RepC_14. II in red indicates unstable. **D**, Predicted XtalPred crystallisability classification using random forest classifier (RF) method for uncleaved His-RepC (left) and His-RepC_14 (right) expressed using pCold I.

Basal level expression in the absence of IPTG was observed for His-RepC_14 but not His-RepC. Average cell pellet yield was 20 mg/mL for cultures expressing His-RepC, and 40 mg/mL for cultures expressing RepC_14. The predicted II value of uncleaved His-RepC_14 (40.9) was lower than the II value for uncleaved His-RepC (44.6). Despite having an improved score, His-RepC_14 was still predicted to be unstable in solution (Fig. 3.19C). No change was observed in the predicted XtalPred crystallisability scores for the two proteins (Fig. 3.19D). Both proteins obtained a score of 10, indicating that the chances of crystallisation were very low.

Expression levels of MBP-RepC and MBP-RepC_14 using pMAL-c2X in T7 Express were compared using SDS-PAGE (Fig. 3.20A, B). Both MBP-RepC and MBP-RepC_14 were expressed in the soluble fraction as evidenced by the presence of a distinct band corresponding to the predicted molecular mass of the two proteins. The apparent intensity of the target protein bands was identical. Average cell pellet yield was identical for cultures expressing either protein (40 mg/mL). The predicted II value of factor Xa-cleaved RepC_14 (39.3) was lower than the II value for factor Xa-cleaved RepC (43.4). As a result, the cleaved RepC_14 protein was predicted to be stable in solution (Fig. 3.20C). The predicted XtalPred crystallisability score for factor Xa-cleaved RepC_14 was also predicted to be improved in comparison to factor Xa-cleaved RepC (score 8 versus 9) (Fig. 3.20D). Despite this improvement, both proteins were predicted to be unlikely to crystallise.



	MBP-RepC_14 (uncleaved)	RepC (factor Xa-cleaved)
MW (kDa)	68.3	25.8
pI	5.2	5.4
II	27.5	39.3

Fig. 3.20 Comparative miniaturised expression trials of MBP-RepC and MBP-RepC_14 proteins. **A**, SDS-PAGE analysis of uninduced (UI) and IPTG-induced cell lysate (CL) and soluble (SOL) fractions extracted from E. coli T7 Express liquid culture transformed with pMAL-c2X-RepC after overexpression at 18 °C for 16 h. Target product band (69.7 kDa) indicated. **B**, SDS-PAGE analysis of uninduced UI and IPTG-induced CL and SOL fractions extracted from E. coli T7 Express liquid culture transformed with pMAL-c2X-RepC_14 (encoding RepC truncated by 14 C-terminal residues), after overexpression at 18 °C for 16 h. Target product band (68.3 kDa) indicated. Non-transformed (NT) CL and SOL fractions included for comparison. M, protein molecular mass marker. **C**, Predicted Expasy ProtParam molecular mass (MM), theoretical pI, and instability index (II) parameters for uncleaved MBP-RepC_14 and cleaved RepC_14. II in green indicates stable, II in red indicates unstable. **D**, Predicted XtalPred crystallisability classification using random forest classifier (RF) method for cleaved RepC (left) and RepC_14 (right) expressed using pMAL-c2X.

3.4 Discussion

3.4.1 In silico analyses of African begomovirus Rep proteins reveals conserved functional motifs typical of other viral Rep proteins.

The present study was aimed at characterising and comparing the domain organisation of the Rep proteins of ToCSV V30 and related African mono-and bipartite begomoviruses and comparing identified sequence motifs with that of distantly related viral Rep proteins. The well-characterised Rep protein sequence of TYLCSV was included for comparison. The DNA binding activity of Rep is mediated by sequence

specific interaction of Rep with cognate repeat sequence (iteron) motifs present in the 5' IR situated in the vicinity of the TATA box core promoter sequence (Argüello-Astorga et al., 1994). The iteron sequences identified for several African begomoviruses were shown to vary greatly in their length, sequence identity and arrangement. This was expected since iteron sequences are known to differ even amongst closely related geminiviruses or strains of a particular species (Chatterji et al., 1999; Fontes et al., 1994; Hanley-Bowdoin et al., 2000; Lazarowitz et al., 1992). This variation allows for specific sequence recognition by Rep to ensure that binding and subsequent replication of a specific genome is mediated by its cognate Rep protein. A comparative study of two strains of ToLCNDV sharing 94% sequence identity revealed that although their iteron sequences only differ by two nt, the Rep of one strain is unable to efficiently bind the iteron of the other strain (Chatterji et al., 1999).

It may therefore seem contradictory that in some cases the iteron sequences of distinct species are identical. The iteron sequence of the exemplar isolate of ToCSV has been previously reported to be identical to that ToLCSDV/Sha with the sequence being reported as GGTAC (Pietersen et al., 2008). The current study expanded the length of the identified iteron to TCGGTACCCA which is identical to the iteron of TbLCZV and nearly identical to that of ToLCSDV/Sha and ToLCKMV (one nt mismatch). Whether the Rep of ToCSV can replicate the DNA of these two other viruses is unknown. Although iteron sequence identity plays a major role in defining replication specificity, it is not the sole determining factor involved in the process (Londoño et al., 2010). The presence of iteron motifs flanking the TATA box promoter on either side has also been implicated in the sequence-specific repression of the transcription of the C1 gene and other viral cs genes by Rep (Eagle et al., 1994; Sunter et al., 1993). This Rep-mediated repression could occur with viruses sharing iteron sequence identity but not arrangement (Argüello-Astorga et al., 1994). It would be worthwhile to determine if ToCSV Rep is able to repress the expression of heterologous Rep proteins by binding to identical iteron sequences found in the genomes of related begomoviruses such as TbLCZV which exist in the same geographic location. The presence of additional distinct *cis*-acting elements in the IR could also determine sequence specificity and warrant further investigation in the case of African begomoviruses as this has implications for their evolution in the field.

A broad region of the Rep N-terminal domain has been implicated in the sequence-specific DNA-binding activity both in vitro and in vivo (Jupin et al., 1995; Orozco et al., 1997, 1998). Multiple sequence analysis of a wide range of geminivirus iteron sequences in a heuristic manner has reduced the size of the presumed iteron-binding region to a stretch of eight to ten N-terminal residues termed the IRD (Argüello-Astorga and Ruiz-Medrano, 2001). Geminiviruses such as ToCSV that display a core iteron sequence of GGTAC contain a predicted IRD motif having the aa sequence PKRFQI (Argüello-Astorga and Ruiz-Medrano, 2001). This aa sequence motif was identified in all three ToCSV variant Rep proteins as well as in the Reps of TbLCZV and ToLCSDV/Sha. The IRD motif of ToLCKMV contained a lysine instead of an arginine at IRD aa position three. The presence of identical Rep IRDs in these three viruses lends further support to the notion that their Rep proteins could modulate each other's replication cycle. It was previously reported that a single residue difference in the IRD in combination with two nt mismatches in the iterons of two strains of ToLCNDV were responsible for the differing mild and severe disease phenotypes (Chatterji et al., 1999). The two wild type Rep proteins were strictly dependent on their respective cognate sequences for successful replication. Mutagenesis to alter the sequence identity of the mild strain to that of the severe strain resulted in a change to severe disease phenotype and vice versa (Chatterji et al., 1999). This suggests that even slight changes to the sequences involved in Rep-binding activity can result in differences in binding specificity with significant consequences for disease development. The differing symptom phenotypes observed in ToCSV V30 (severe) and V22 (mild) cannot be attributed to this phenomenon since both variants have identical iteron and Rep IRD sequences. An additional specificity determinant (SPD) element comprised of two discrete residues implicated in sequence-specific binding and replication has been identified a few residues downstream of N-terminal domain motif II (Londoño et al., 2010). In both ToCSV variant sequences the SPD motif (Asn69, Lys71) was shown to be identical to that of TbLCZV but not ToLCSDV/Sha (Lys69, Gln71) or ToLCKMV (Arg70, Thr72). It would be interesting to conduct mutagenesis to determine if modification of the SPD results in altered Rep binding, and whether heterologous binding of ToCSV Rep to ToLCSDV/Sha or ToLCKMV is impaired by the differing SPD motifs.

The endonuclease domain of Rep (ToCSV Rep aa 1-120) is responsible for the DNA cleavage activity of geminivirus Rep (Orozco et al., 1997). Three functional motifs implicated in this activity have been described (Ilyna and Koonin, 1992; Koonin and Ilyna, 1992; Orozco and Hanley-Bowdoin, 1998). Using alanine substitution, important residues present in motif I (Phe16, Tyr19), motif II (His58, L59, His60), and motif III (Tyr104) of TGMV Rep were identified. Mutation abolished the dsDNA binding, cleavage and ligation activities of Rep with DNA replication being impaired as a result (Orozco and Hanley-Bowdoin, 1998). In the present study, the phenylalanine of motif I was highly conserved whilst the tyrosine of motif I was found in 18/23 African begomovirus Rep proteins. In the remaining Reps a phenylalanine (also aromatic) was present at this position. A conserved threonine was also identified at the third residue position of motif I. High sequence conservation of these residues has been shown in Rep proteins of other geminiviruses and distantly related virus Rep proteins involved in RCA (Nash et al., 2011; Vadivukarasi et al., 2007). The HLH sequence of motif II was conserved in 21/23 African begomovirus Rep proteins whilst two Rep sequences contained an isoleucine (also hydrophobic) between the two histidine residues. The presence of the histidine residues of motif II separated by a hydrophobic residue classifies these Rep proteins as HUH endonucleases (Kazlauskas et al., 2019; Tarasova and Khayat, 2022; Venkataraman and Selvarajan, 2019). The tyrosine of motif III was conserved in 23/23 Rep sequences. The catalytic tyrosine of motif III and lysine at motif III aa position four is highly conserved in the Rep proteins of geminiviruses and other ssDNA viruses including circoviruses and nanoviruses (Tarasova and Khayat, 2022). The sequence analysis conducted herein confirms that that ToCSV and related African begomovirus Rep proteins contain the functional motifs necessary for the predicted catalytic function of the endonuclease domain.

The GRS motif was identified as an additional Rep sequence motif containing several residues necessary for endonuclease function. The consensus GRS sequence 1-RXFDXXXXXXXXFHPNIQXAKX-23 has been shown to be conserved amongst diverse geminivirus genera (Nash et al., 2011). GRS sequence position X indicates consensus residue not conserved amongst diverse genera. Sequence comparison of ToCSV and African begomovirus Reps with the GRS consensus sequence showed matching consensus sequence except for isoleucine at position 18 (present in 20/23 Rep proteins). African begomovirus Rep proteins thus display the highly conserved

GRS motif residues found in other geminiviruses and therefore this region of V30 Rep is implicated in GRS-mediated endonuclease activity.

Homology modelling of the N-terminal domain of V30 Rep using the solved structure of TYLCSV Rep (PDB ID 1L5I) as the template was conducted to visualise and compare the arrangement of the important endonuclease sequence motifs. Topology mapping served to provide a useful 2-dimensional visual comparison of homology models with solved ssDNA virus Rep structures. Two monomeric homology models were generated, and structure quality assessment found the RepN-SM structure to be more energetically favourable and so this structure will be discussed herein. Two residues of the V30 Rep IRD (Ile9 and Gln10) and Asn69 of the SPD motif were closely associated and formed part of two β-strands of an antiparallel β-sheet. The side chains of these residues were directed towards the solvent-exposed surface suggesting a possible region for DNA contact to occur. This compares favourably with published homology models of the Rep protein of two strains of TYLCSV which display an identical ss feature (Londoño et al., 2010). The positions of key IRD and SPD motif residues (as per Londoño et al., 2010) were mapped to the solved structures of other virus Reps (data not shown). It was found that in both WDV and TYLCSV structures (Campos-Olivas et al., 2002; Everett et al., 2019), these key residues were located on the curved N- and C-termini of extended β-strands containing motif I and II respectively. These residues were also shown to be solvent-exposed providing evidence for the role of this region in DNA binding. The observed ss of the modelled V30 Rep DNA binding-associated residues (Ile9, Gln10 and Asn69) agreed with the PSIPRED-predicted consensus ss of these regions obtained for the Rep proteins of African begomoviruses.

The arrangement of key residues belonging to motif I, motif II and motif III was consistent with that of solved structures of other viral Rep N-terminal domains (Campos-Olivas et al., 2002; Everett et al., 2019; Vega-Rocha et al., 2007a, 2007b). V30 Rep residues belonging to motif I and motif II were predominantly located on the first and third β-strands of an antiparallel β-sheet respectively. The catalytic tyrosine of motif III was located on a helix positioned above the β-sheet. The proximity of the key residues of these motifs illustrated using liquorice representation suggests the presence of a catalytic site in this region of RepN-SM. The structures of WDV and PCV2 Rep N-terminal domains complexed with their respective ssDNA

nonanucleotide sequences (containing Ori), and the manganese ion (Mn^{2+}) cofactor have been described (Tompkins et al., 2021). The authors showed that key residues present in the three motifs engage in protein-DNA contact. In addition, the two histidine residues of WDV motif II and the His57 and Gln59 of PCV2 motif II are involved in coordinating of Mn^{2+} at the cleavage site (Tompkins et al., 2021). Further V30 Rep N-terminal domain homology modelling in conjunction with molecular docking of a ssDNA template is necessary to determine whether the catalytic residues present in the putative active site adopt a similar conformation to that present in the WDV Rep structure. The PSIPRED-predicted consensus ss for African begomovirus Rep proteins agreed with that observed for the homology model of V30 Rep endonuclease domain, suggesting that these proteins are likely to adopt similar tertiary structure.

In WDV and TYLCSV Rep, most of the GRS motif residues are located on a β-hairpin loop containing two antiparallel β-strands (Campos-Olivas et al., 2002; Everett et al., 2019) although the β-hairpin structure was not described in the published report of the WDV structure (Everett et al., 2019). Surprisingly, these β-strands are absent in the RepN-SM homology model although the long loop structure is present. The predicted ss for this region of the GRS in African begomovirus Rep proteins did not contain predicted consensus strands or helices which implies that this region may have a variable ss. In the RepN-SM homology model, GRS motif residues 17 to 20 (Asn90, Ile91, Gln92, Gly93) form β4 of the antiparallel β-sheet adjacent to strands containing key residues implicated in endonuclease activity. The corresponding GRS motif residues of TYLCSV (Asn90, Ile91, Gln92, Gly93) and WDV (Asn93, Ile94, Gln95, Ala96) also form β4 in their respective structures (Campos-Olivas et al., 2002; Everett et al., 2019). A novel “ssDNA bridging motif” (sDBM) was recently described encompassing the ten C-terminal aa of the GRS (containing β4). The sDBM of WDV has been shown to bind the ssDNA nonanucleotide via sequence-specific protein-ssDNA contacts and stabilise the nonanucleotide structure to adopt a conformation that facilitates cleavage (Everett et al., 2019). Although PCV2 Rep does not contain a GRS motif, Tompkins et al. (2021) identified the corresponding PCV2 sDBM motif sequence and confirmed that residues present in this region serve the same function described for geminivirus Rep sDBM.

The oligomerisation domain was not modelled for V30 Rep (aa 121 to 180) since no homologues with sufficient sequence identity were identified in the PDB. The

sequence of the oligomerisation domain is not conserved amongst SF3 helicases (Clérot and Bernardi, 2006). The oligomerisation domain of African begomovirus Reps was predicted to contain four consensus helices at high confidence. This PSIPRED prediction correlates with the observed ss solved structures of other DNA virus Reps. The monomeric units of the oligomerisation domains of PCV2, BPV and AAV2 Reps each contain four helices arranged in a small bundle. In oligomeric form, these bundles are organised into a rigid collar-like structure that is situated adjacent to the helicase domain oligomeric ring structure (Enemark and Joshua-Tor, 2006; James et al., 2003; Sanders et al., 2007; Tarasova et al., 2021). It is anticipated that the oligomerisation domain of begomovirus Rep forms a similar collar-like structure, since oligomerisation is required for helicase activity in both geminivirus and other viral helicases (Choudhury et al., 2006; James et al., 2003; Tarasova et al., 2021). The N-terminal oligomerisation domain sequence RSARGG (V30 aa positions 121 to 126) was shown to be conserved in 20/23 African begomovirus Rep proteins. Additionally, this region was predicted to contain a series of disordered residues in some of the Rep sequences analysed including V30 Rep. This sequence has previously been shown to be highly conserved amongst begomoviruses and it was demonstrated that the RGG sequence of TYLCSV Rep (Rep aa 124 to 126) is involved in regulating the transcription of its own C1 ORF (Sardo et al., 2011). This region also contains residues implicated in ATPase/helicase activity (Choudhury et al., 2006). In context of the predicted disorder, this could involve interaction of this region with viral and/or host factors, although the exact mechanism by which this short region is able to execute this function is yet to be described. Since the oligomerisation domain of geminivirus Rep is known to facilitate binding of Rep to itself as well as several other viral and host proteins (Krenz et al., 2011; Rizvi et al., 2015), the role that these specific residues could play in Rep-protein binding should be investigated further.

The ATPase/helicase domain of geminivirus Rep (ToCSV Rep aa 181-330) is responsible for the ATP hydrolysis and coordinated DNA unwinding (helicase) activity during viral genome replication (Choudhury et al., 2006; Clérot and Bernardi, 2006; Desbiez et al., 1995; Hickman and Dyda, 2005). Four functional motifs which are necessary for this activity are present in all SF3 helicases (Hickman and Dyda, 2005). A lysine residue in WA motif and an aspartate in WB motif of tomato leaf curl New Delhi (ToLCNDV) and ToLCGuV Rep proteins were shown to be indispensable for

helicase activity in vitro and viral replication in vivo (George et al., 2014; Ruhel et al., 2021). The lysine in WA is also critical for helicase activity in TYLCSV Rep (Clérot and Bernardi, 2006). These two residues were 100% conserved in the WA and WB motifs of V30 Rep and other African begomovirus Rep proteins (V30 Rep Lys227, Asp261). A conserved asparagine was also found in the C motif of V30 and other Rep proteins and this residue is known to be conserved in the C motif of other viral SF3 helicases as well (Gai et al., 2004; James et al., 2003; Sanders et al., 2007; Tarasova et al., 2021). The B' motif is unique to SF3 family helicases and contains either a lysine or arginine at the first sequence position whilst a lysine at B' motif aa position 15 is known to be highly conserved amongst all SF3 helicases (James et al., 2003, 2004; George et al., 2014; Yoon-Robarts et al., 2004). Lysine residues at both positions were identified in the B' motif of all begomovirus Rep proteins analysed in this study. The confirmed presence of sequences corresponding to important helicase domain functional motifs supports the classification of these African begomovirus Rep proteins as SF3 helicases. One key feature missing in previously described geminivirus Rep proteins is the arginine finger motif located downstream of the C motif of other DNA virus Rep proteins including BPV E1, PCV2 Rep, SV40 Tag, and AAV2 Rep (Clérot and Bernardi, 2006; George et al., 2014; Hickman and Dyda, 2005; James et al., 2003; Sanders et al., 2007; Tarasova et al., 2021). Interestingly, in the current analysis, 4/23 African begomovirus Rep proteins were found to contain an arginine residue downstream of the C motif. Whether this apparently rare arginine functions as an arginine finger motif in these selected Rep proteins requires further study. Several conserved lysine residues are present in this region in African begomovirus Reps. Despite both having a positive charge, lysine and arginine are significantly different in terms of their chemical features and lysine typically does not function as an efficient substitute for arginine in the arginine finger motif (Yukawa et al., 2015). How geminivirus Rep is able to execute ATP hydrolysis without an arginine finger is an intriguing mystery which will likely remain so until a high-resolution structure of oligomeric geminivirus Rep is solved. The C-terminal domain spanning the 29 C-terminal residues of Rep is discussed further in Section 3.4.2.

Homology modelling of the partial helicase and C-terminal domains of V30 Rep was done. Unfortunately, no solved monomeric or oligomeric structure of any geminivirus Rep helicase domain exists in the PDB and the most relevant template identified was

the solved hexameric structure of BPV E1 (PDB ID 2GXA). A monomer (RepC-P) and a hexamer (RepC-SM) were obtained. The N- and C-termini of RepC-P were modelled via ab initio methods which is known to provide unreliable results (Kelley et al., 2015). Topology mapping was conducted as for RepN models to evaluate and compare of the arrangement of important helicase functional motifs in V30 Rep with that of other SF3 helicases. Structure quality assessment showed that although the overall quality of both structures was low, the RepC-SM hexamer scored better than the monomer. The ss features corresponding to β1 to β4 (parallel β-sheet) and the first helix obtained higher local quality estimate scores indicating that these regions were modelled more favourably. Although many SF3 helicases generally exist as functional hexameric structures, AAV2 Rep is known to form heptamers and a heptamer/hexamer mixture upon binding ATPγS (Santosh et al., 2020). More importantly, the Rep proteins of begomoviruses have been shown to form oligomers ranging from hexamers up to dodecamers and even higher order oligomers including 24-mer structures (Choudhury et al., 2006; Clérot and Bernardi et al., 2006; George et al., 2014). It is critical to note that the oligomeric conformation required for helicase activity was determined to be at least dodecameric for TYLCSV Rep (Clérot and Bernardi, 2006) and approximately 24-meric for MYMIV Rep (Choudhury et al., 2006). This implies that the biologically active form of begomovirus Rep may display a unique arrangement that could render at least part of the modelled hexameric form obsolete once the molecular structure is solved in future. Nevertheless, the hexameric V30 RepC-SM model is described herein.

The structure of the monomeric units of RepC-SM was consistent with that of other SF3 helicases including BPV E1 (Enemark and Joshua-Tor., 2006; Hickman and Dyda, 2005) containing a parallel β-sheet with a few helices located above and below it. Instead of containing five β-strands however, the β-sheet contained four β-strands, missing the second N-terminal β-strand present in other SF3 helicase structures. The arrangement of key helicase domain motifs was consistent with that of the published structures of other SF3 helicases (Enemark and Joshua-Tor., 2006; Gai et al., 2004; Santosh et al., 2020; Tarasova et al., 2021). Despite the somewhat rudimental fold of the RepC-SM monomer, the V30 Rep structure displayed the hallmarks of an SF3 helicase. The WA motif was located to β1 up to the C-terminal end of the first helix. WB and C motif were mapped to two β-strands adjacent to β1. Important RepC-SM

WA (Gly221, Gly226, Lys227, Thr228), WB (D261, D262), and C motif (Asn304) residues were shown using liquorice representation and were found to closely associate together, forming a probable catalytic site. In SV40 Tag, WA lysine and threonine, WB aspartate, and C motif asparagine residues interact with the phosphate groups of ATP (Gai et al., 2004). Residues on the adjacent (trans-) monomer that coordinate with WA and WB on the cis-monomer include the arginine finger as well as selected B' motif residues (Enemark and Joshua-Tor, 2006; Gai et al., 2004; Tarasova and Khayat, 2022). The close association of the WA, WB and C-motif residues in RepC-SM suggest a similar ATP binding site. Molecular docking of RepC-SM with ATP/ATP analogues could provide more mechanistic details of this interaction. The PSIPRED-predicted consensus ss of regions spanning these motifs in African begomovirus Reps corresponded with that observed for the V30 Rep homology model suggesting high structure similarity amongst different begomovirus Rep proteins.

The lysine present at B' motif position 15 is conserved in SF3 helicases and has been shown to be essential for helicase activity in AAV2 Rep (Yoon-Robarts et al., 2004). Using alanine substitution, the role of selected B' motif and β-hairpin loop aa (Lys272, Arg279, Asp280, Lys286, Lys289 and Pro290) in the activities of ToLCNDV and ToLCGV Rep in vitro has been examined (George et al., 2014; Ruhel et al., 2021). These residues are not required for ATP hydrolysis, but Lys272 and Lys289 are required for DNA binding and helicase activity. Interestingly the P290A mutant did not display impaired DNA binding but helicase activity was abolished. Mutant Rep proteins lacking Arg279, Asp280, and Lys286 displayed significantly reduced helicase activity (George et al., 2014; Ruhel et al., 2021). In V30 RepC-SM the B' motif is partially located to a β-hairpin loop which extends into the central channel of the hexamer. This positioning confirms that like ToLCNDV and ToLCGV Rep, this region of V30 Rep is likely also involved in the binding and unwinding of DNA. In addition, the sequence identity of these key residues are identical for the three Rep proteins.

Structural data for other SF3 helicases demonstrates that this region is highly dynamic and is integral in coupling of ATP hydrolysis with DNA binding to coordinate helicase activity (James et al., 2004; Yoon-Robarts et al., 2004). In the current study, the key residues identified by George et al. (2014) and Ruhel et al. (2021) were mapped to the V30 RepC-SM structure. The location of RepC-SM Lys286, Lys289 and Pro290 on the loop present between the two β-strands corresponded with that observed for modelled

structures of ToLCNDV and ToLCGV Reps (George et al., 2014; Ruhel et al., 2021). Homology modelling carried out in this study was useful to determine the predicted positions of key functional motifs present in the N- and C-terminal domains of Rep. Additional structure refinement is necessary before more detailed molecular modelling studies of ToCSV Rep are undertaken. Such studies will serve to improve the understanding of the molecular intricacies that define an evidently fine-tuned and highly coordinated enzymatic process necessary for viral replication.

3.4.2 The disordered N- and C-termini of V30 Rep warrant increased scrutiny for their role in the functionality of Rep.

The C1 ORFs of V30, V22 and the exemplar isolate of ToCSV (AF261885) were predicted to contain a 5' extension encoding a putative N-terminal Rep protein segment 12 aa in length. Putative 5' ORF extensions were also identified in the C1 ORFs of eight other African monopartite begomovirus isolates. The C1 of ToCSV and these other begomovirus isolates all contained the "bona-fide" C1 start codon encoding the canonical Rep aa 1 (methionine) located 14 to 15 residues upstream of the conserved phenylalanine of Rep endonuclease domain motif I. This meant that canonical Rep was encoded regardless of whether these 5' extensions were spurious or not. The existence of C1 ORF 5' extensions has not been addressed in published literature although the annotated GenBank Rep protein sequence descriptions of Datura leaf curl virus (DaLCV: MF402918), tomato leaf curl Seychelles virus (ToLCSCV: AM491778), ToLCKMV (AJ865340), ToLCToV (AM701768), and tomato leaf curl Anjouan virus (ToLCAnV: AM701758) each display the predicted Rep N-terminal extensions. The references for these sequences do not mention these unusual Rep sequence extensions (Delatte et al., 2005; Lefeuvre et al., 2007a, 2007b; Mohammed et al., 2018). Isolates belonging to two distinct Mozambican ToCSV variant groups (ToCSV-Chok I and ToCSV-Chok-II) (Sande et al., 2021) were analysed in the current study using ORF finder. It was found that isolates belonging to the Chok-I group were predicted to encode the 12 aa Rep N-terminal extension, whereas those in Chok-II encoded canonical Rep without the extension (data not shown). Whether this is merely a coincidental artefact of in silico translation of two divergent sequence clusters, or a biologically relevant feature resulting from their

evolution is up for debate. More sequence information is required to determine if this 5' extension is observed in the C1 ORF of ToCSV variant/strain groups present elsewhere in Southern Africa.

Transcript mapping of the DNA-A of several begomoviruses showed that the C1 ORFs of the begomovirus sequences analysed did not contain putative C1 ORF 5' extensions and the bona-fide start codon was recognised as the true C1 ORF start (Akbar et al., 2012; Mullineaux et al., 1993; Shivaprasad et al., 2005). It is possible that these seemingly rare C1 ORF 5' extensions may represent spurious artefacts. Transcript mapping of CLCuBuV however, revealed the presence of an atypical C4 ORF (Akbar et al., 2012). In most begomoviruses the small C4 ORF (~258 nt) is nested within the C1 ORF (Zerbini et al., 2017). In CLCuBuV, C4 is longer than normal (181 aa), and the C4 ORF start codon is located upstream of both the C1 ORF start codon as well as the iteron motif-associated TATA box promoter. Additionally, two in-frame start codons are present within the C4 ORF which could result in two shorter protein products (Akbar et al., 2012). Due to the confirmed presence of a long (~1.7 kb) transcript encompassing the long C4 ORF, the authors speculated that multiple proteins including the longer C4 may be translated and that additional experiments were necessary to determine whether this was the case. Since begomovirus genomes are known to be a source of long RNA transcripts (Pooggin, 2013) it is possible that multiple products could be translated from ORFs containing 5' terminal extensions albeit at a lower frequency than canonical proteins. In the present study, a transcript spanning the putative C1 ORF 5' segment into the canonical C1 region was detected using RT-PCR with sequence-specific primers. Additional experiments are necessary to determine the true length of the C1 ORF transcript and whether this extended Rep protein is translated.

If both canonical and N-terminally extended Rep (Rep-N) proteins are translated, it is reasonable to argue that these two proteins would likely exhibit differing activities that could benefit the virus. Comparative analysis of the predicted disorder for V30 Rep and Rep-N showed that the 12 aa N-terminal segment of Rep-N was predicted to be disordered with protein-binding potential. Rep-N aa 13 to 22 contains the IRD motif involved in sequence-specific binding of Rep to the IR (Argüello-Astorga and Ruiz-Medrano, 2001). Due to their proximity, it is possible that the N-terminal segment present in Rep-N may alter the functionality of the IRD by altering its binding affinity to

DNA and/or other proteins. Comparative binding assays using Rep and Rep-N should be conducted to investigate this possibility. The predicted disordered region with protein-binding potential present in the oligomerisation domain of Rep was lost in Rep-N which may have consequences for the functioning of this domain since it is known to be involved in protein-protein binding (Rizvi et al., 2015; Sardo et al., 2011). Prediction of intrinsic protein disorder can provide useful clues as to which proteins/protein regions may potentially behave in a manner that could ultimately serve to regulate their biological activities (Jones and Cozzetto, 2015). Such predictions require experimental validation and conclusions cannot be drawn from this predictive data alone. This is especially true for proteins such as Rep-N which contain putative regions that have not yet been verified to exist. Nevertheless, such unique sequence features provide an excellent opportunity to gain a better understanding of the more "neglected" regions of the Rep protein such as the oligomerisation domain.

Another begomovirus Rep domain region for which little published information exists is the short C-terminal domain spanning V30 Rep aa 331 to 259. This domain was first annotated in TYLCSV Rep (Clérot and Bernardi, 2006). The authors did not elaborate on the parameters used to delineate the N-terminal boundary of this domain, but the boundary positions for TYLCSV Rep (aa 331 to 359) were applied when identifying C-terminal domains in African begomovirus Reps. The precise function of this domain is unknown (Clérot and Bernardi, 2006). Removal of 39 aa from the C-terminal end of the Rep of TGMV resulted in a loss of viral DNA replication in vivo (Orozco et al., 1997). To investigate the region deleted by Orozco et al. (1997), the aa sequence of TGMV Rep was aligned with that of TYLCSV and it was found that the C-terminal domain of TGMV is only 18 aa long (data not shown). The remaining region deleted by the authors included 21 aa of the helicase domain C-terminus. It is possible that removal of this portion of the helicase domain was responsible for the observed loss of replicative ability rather than due to the loss of the C-terminal domainas implied by Clérot and Bernardi (2006). James et al. (2003) reported that removal of the 46 C-terminal residues of AAV2 Rep did not alter its helicase activity in vitro in comparison to full length Rep protein.

The predicted consensus β-strand of the C-terminal domain of begomovirus Rep (V30 Rep aa 332 to 335) was mapped to the final β-strand forming the parallel β-sheet of the putative catalytic ATPase site in both RepC-P and RepC-SM homology models.

This final β-strand was located adjacent to the β1 containing partial WA motif in the homology models as well as in the helicase domain of the other SF3 helicase structures (Gai et al., 2004; James et al., 2003; Sanders et al., 2007; Tarasova et al., 2021). This implies that the demarcation of the N-terminal boundary of the begomovirus Rep C-terminal needs to be revised. The final β-strand is too intimately linked to the remaining helicase domain core β-strands to be assigned to a separate domain. One characteristic feature of the C-terminal domain shared by all the begomovirus Rep proteins analysed is the presence of a C-terminal disordered region (10-20 aa in length) predicted to have protein-binding potential. This disordered region contained two to six acidic residues (Asp/Glu). The predicted disorder of other SF3 helicases was compared with begomovirus Rep proteins (data not shown). It was discovered that BPV E1 helicase was also predicted to contain a C-terminal disordered region (27 aa) containing ten acidic residues. This region of BPV E1 has been previously termed the C-terminal acidic tail (C-tT) and is known to be a highly flexible region as a result of its intrinsic disorder (Chaban et al., 2015; Whelan et al., 2012). The C-tT has been implicated in stabilising the hexameric structure of BPV E1 to ensure processivity of the helicase activity and is required for replication activity in vivo (Whelan et al., 2012). Incredibly, the structure of the BPV E1 C-tT region has recently been defined using cryo-EM (Javed et al., 2021) and each monomeric subunit was shown to exist as a loop extending from the helicase domain into cleft regions that exist between adjacent oligomerisation domain subunits (Javed et al., 2021). Such a region has not been described for any geminivirus Rep protein, but acidic tails are known to exist in other helicases including SV4 Tag and bacteriophage T7 helicase-primase (Sawaya et al., 1999; Javed et al., 2021). The role (if any) that this putative acidic tail region plays in helicase activity of V30 Rep should be investigated further.

Two C-terminal domain deletion mutants of ToCSV V30 were obtained: a 14 aa deletion mutant (V30ΔC1-d(14) and a 22 aa deletion mutant (V30ΔC1-d(22). Both mutant Rep proteins were predicted to have lost their C-terminal disorder and in the case of V30ΔC1-d(14), a significant reduction in predicted oligomerisation domain disorder. Infectivity assays revealed that plants infected with the V30ΔC1-d(14) mutant had significantly reduced viral load whilst V30ΔC1-d(22) was not able to infect *N. benthamiana* despite repeated inoculation attempts. It can be concluded from these observations that removal of the 14 C-terminal residues encompassing the disordered

region does not completely abolish infectivity and that the mutant Rep protein must retain some level of helicase activity which is necessary for successful viral replication in vivo (Choudhury et al., 2006; George et al., 2014). Whether the delay in symptom expression is due to reduced Rep enzymatic activity as a consequence of structural changes or due to disrupted/altered Rep protein-protein interactions remains to be determined. The inability of V30ΔC1-d(22) to successfully infect *N. benthamiana* suggests that the residues located at Rep aa positions 338 to 345 are of substantial importance in the normal functioning of Rep. This does not necessarily mean that Rep ATPase/helicase activity was abolished in this mutant. It is possible that truncation of 22 C-terminal residues resulted in disruption of important protein-protein interactions necessary for infection instead of abolishing ATPase/helicase activity by disrupting the native structure of Rep. In vitro enzymatic assays using C-terminally truncated or C-terminal domain mutant Rep proteins will provide greater insight into the role that these residues may contribute to the key enzymatic activities of Rep that are necessary for successful infection.

3.4.3 Expression and purification of recombinantly expressed V30 RepC proteins – a complex puzzle with many pieces.

Structural studies of a particular protein undertaken to reveal the intricacies of their enzymatic activities at the atomic level, or to probe their interactions with other molecules requires sufficient quantities of a pure target protein. In the current study, recombinant expression of ToCSV V30 Rep in *E. coli* was done. This well-studied host is favoured since it is readily transformed with various expression vectors, can be grown rapidly using simple media to produce high quantities of target proteins (Rosano and Ceccarelli, 2014). Several *E. coli* strains are available with differing properties that are advantageous to recombinant protein expression (Sørensen and Mortensen, 2005). The BL21(DE3) strain has been used previously in other studies of partial/full-length geminivirus Rep proteins (Choudhury et al., 2006; George et al., 2014; Kushwaha et al., 2017; Nash et al., 2011; Ruhel et al., 2021). The reported difficulties experienced in transformation and expression of pCold I-RepC in the present study suggested that this modified expression vector or the recombinant protein product was inherently toxic to these strains.

All three *E. coli* strains used in this study employ an expression system using bacteriophage T7 RNA polymerase (T7RNAP) to yield high quantities of recombinant protein (Angius et al., 2018; Studier and Moffatt, 1986). The λDE3 lysogenic prophage harbours the gene encoding T7RNAP. Expression of T7RNAP is under control of the *lac*UV5 promoter element with induction using IPTG although BL21(DE3) has been reported to have some level of basal T7RNAP expression which results in "leaky" expression of target protein-encoding genes (Samuelson, 2011). In cases where leaky expression of toxic proteins occurs, cell growth may be arrested early prior to induction due to the recombinant proteins interfering with normal host cell processes. In addition, the selection of cells displaying lower expression levels may occur (Pan and Malcolm, 2000; Samuelson, 2011). The pLysS plasmid (encoding T7 lysozyme to decrease levels of basally expressed T7RNAP) ensures stricter regulatory control to minimise basal expression of toxic proteins which could be detrimental to cell growth prior to induction (Francis and Page, 2010; Miroux and Walker, 1996; Pan and Malcolm, 2000). BL21(DE3)pLysS is often used to express proteins exhibiting toxicity in BL21(DE3) (Hoffman et al., 1995; Francis and Page, 2010; Samuelson, 2011). The impaired growth and lack of recombinant expression seen in BL21(DE3)pLysS when transformed with pCold I-RepC suggests that even this stricter level of expression regulation was insufficient to negate the inherent toxicity of His-RepC expressed using pCold I. Expression of His-TF-RepC was successful using BL21(DE3)pLysS which suggests that the presence of the His-TF tag may have reduced the inherent cytotoxicity of Rep. The pCold TF vector has been previously used to improve the expression and solubility of other proteins with the N-terminal His-TF tag presumably facilitating correct folding and shielding regions of the recombinantly expressed protein which would otherwise interfere with the host (Hu et al., 2009; Saini et al., 2014). This strongly suggests that expression of V30 Rep with a large chaperone tag may be necessary in future to inhibit the cytotoxic effects of untagged Rep in BL21 strains of *E. coli*.

His-RepC was obtained in the soluble fraction using the T7 Express strain which contains the T7RNAP gene in the chromosomal *lac* operon ensuring significantly less basal expression than that seen in strains containing λDE3. In addition, this strain is more suited to expression of proteins that interact with DNA and has greater stress tolerance (Fomenkov et al., 2017; Samuelson, 2011). The pCold vector series utilises

the *E. coli* cold shock protein A (*cspA*) promoter system (induction via low temperature) which is also prone to leaky expression (Francis and Page, 2010). To mitigate this issue, the creators of pCold expression vector series have included the lac operon to prevent basal expression prior to the addition of IPTG at low temperature (Qing et al., 2004). Whilst T7 Express was successfully used to express His-RepC in the soluble fraction, future expression trials of RepC could also be conducted using other *E. coli* strains that have been developed for expression of highly toxic proteins such as *E. coli* C41(DE3) and C43(DE3) (Dumon-Seignovert et al., 2004; Miroux and Walker, 1996). This could be done with the aim of improving yield of soluble protein.

The pMAL-c2X vector has been used previously to express partial begomovirus Rep proteins (Clérot and Bernardi, 2006; George et al., 2014). This vector allows for *tac* promoter-driven expression of high quantities of MBP-tagged protein in the cytoplasm (Riggs, 2000). In the current study, the pMAL-c2X vector yielded consistent levels of soluble MBP-RepC using T7 Express although the oligomeric state of MBP-RepC was not determined. Clérot and Bernardi (2006) reported that partial TYLCSV Rep (aa 120-359) expressed as an MBP fusion formed hexamers in solution and that removal of MBP was required for the formation of higher order oligomers that displayed helicase activity. Higher order oligomerisation has also been reported for other begomovirus Rep proteins showing helicase activity (Choudhury et al., 2006; George et al., 2014). These studies suggest that tag removal (at least in the case of MBP) is required for Rep activity in vitro. On the contrary, His- and glutathione S-transferase (GST)-tagged MYMIV Rep was shown to have ATPase/helicase activity in vitro and the formation of functional higher order oligomers was apparently not impeded by the presence of these tags (Choudhury et al., 2006). This may in part explain the proposed cytotoxicity of V30 His-RepC expressed in BL21 strains using pCold I. Basal expression of biologically active His-RepC (which would form higher order oligomers unhindered) may have resulted in host cell death by inducing DNA damage via direct (DNA-binding) and indirect (host-Rep protein interaction) mechanisms. Additional experiments are necessary to determine the oligomeric states of the recombinant RepC proteins expressed in this study which could provide clues to their activity as tagged fusions.

Removal of both MBP and His-TF tags was successful using Factor Xa and thrombin respectively. The cleaved RepC as well as uncleaved His-RepC were all predicted to be unstable in solution due to their II values (≥40). The statistical methods used to

predict the II value of a protein rely on their primary structure (aa sequence) and does not consider their higher-order structural features (Guruprasad et al., 1990). An additional factor to consider is the surrounding environment which is known to affect the stability of a protein (Gamage et al., 2019) and therefore these predicted values should not be regarded as absolute indicators of instability. In the present study the formation of white precipitate after tag cleavage and/or sample dialysis was observed and was associated with decreased cleaved RepC band intensity on SDS-PAGE suggesting that cleaved proteins were prone to aggregation. Such precipitation was not observed in purified uncleaved MBP-Rep and His-TF-RepC stored on ice at 4 °C for several days. Additional studies to determine the stability of recombinantly-expressed RepC under various conditions including a range of storage temperatures and buffer compositions are required to better understand the conditions which reduce the propensity of the protein to precipitate.

Purification of cleaved MBP-RepC using DEAE anion-exchange chromatography showed that V30 RepC eluted at a NaCl concentration comparable to that reported for DEAE chromatography purification of cleaved TYLCSV Rep (Clérot and Bernardi, 2006). Additional contaminants were reported in the present study (including uncleaved MBP-RepC) indicating that further optimisation of tag cleavage and ion exchange chromatography is required to improve purity of fractions containing cleaved RepC. Despite successful cleavage of His-TF-RepC using thrombin, RepC was found to associate with His-TF as evidenced by co-elution during the second IMAC purification. Despite attempts to disrupt this interaction using 20 mM CHAPS, the two proteins continued to associate (data not shown). Use of alternative affinity tags at either the N- or C- termini should be considered in future expression experiments as several options have been previously described (Kimple et al., 2013; Lichty et al., 2005; Zhao et al., 2013).

Even in the absence of large tags, purification of recombinantly expressed His-RepC was also problematic. Results of IMAC purification using gravity flow revealed that His-RepC did not bind the column despite attempts at optimising buffer conditions. Given that His-RepC may be able to form higher order oligomers in solution, it is possible that the His-tag was positioned in a manner that obscured it and prevented interaction with the immobilised Ni^{2+} present in the resin (Kielkopf et al., 2020). It is possible to improve accessibility by moving the His tag to the C-terminus where it might be more

exposed to the solvent, or by extending the His tag using either a linker sequence or additional histidine residues (Kielkopf et al., 2020). AS precipitation is an alternative purification method that can be used to purify proteins with tags that may be hidden (Duong-Ly and Gabelli, 2014a). In the present study, AS purification was shown to improve the purity of soluble fraction containing RepC (precipitation at 35 % AS) which may prove useful as an inexpensive preliminary purification step prior to subsequent purification using column chromatography methods (Duong-Ly and Gabelli, 2014a; Wingfield, 1998). Gel filtration chromatography was also shown to enhance the apparent purity of His-RepC present in the soluble fraction. The fractionation range of Sepharose 6B is reported to be 10-4,000 kDa (Ó'Fágáin et al., 2011). Gel filtration was conducted in this study as a preliminary experiment and conclusions regarding the molecular weight of His-RepC could not be drawn. Column testing with a high molecular weight marker such as blue dextran 2000 to determine the column void volume, and construction of a suitable calibration curve using several standard proteins of known molecular weight is necessary for further interpretation (Ó'Fágáin et al., 2011; Stanton, 2004). This will assist to determine if the early elution of His-RepC was due to the presence of higher order oligomers or protein aggregation. Elution of His-RepC at the void volume would be indicative of aggregation which is highly unfavourable (Duong-Ly and Gabelli, 2014b).

It is obvious that recombinant expression of V30 RepC requires further optimisation at various stages of expression and purification. The evident cytotoxicity and instability of V30 RepC makes this protein a challenging target for recombinant expression. Although the characterisation of RepC was hampered by the setbacks experienced over the course of this study it is hoped that these preliminary results will guide future expression studies of begomovirus Rep proteins.

3.4.4 Miniaturised expression of truncated RepC proteins – a practical approach to improve yield and solubility.

Infectivity trials in *N. benthamiana* with V30 Rep deletion mutants (Section 3.4.2) were conducted to investigate the role of the disordered C-terminal region in infectivity and symptom development. Truncation of the C-terminal domain of Rep by 14 residues did not completely abolish Rep functionality in vivo suggesting that this region is not essential for the ability of Rep to perform key enzymatic activities required for infection.

Since the V30ΔC1-d(14) mutant was able to infect N. benthamiana, this indicates that some level of native molecular structure required for these activities was preserved in the Rep of this mutant. The results of in vivo infectivity trials using C1 deletion mutants proved useful as a guide to modify the recombinant V30 Rep expression vectors. This was done with the aim of improving the predicted stability of recombinantly expressed RepC by reducing the amount of disordered residues present at the C-terminal end presumably without abolishing enzymatic activity of the helicase domain. The intrinsically disordered region spanning the 14 C-terminal residues of RepC was selected for deletion because it likely contributes to the in vitro instability of the protein. Targeted removal of disordered terminal regions of proteins can be used to increase their stability to make them more amenable to structural studies (Deller et al., 2016; Dosztányi et al., 2007; Ferron et al., 2006; Kamarudin et al., 2014). Large amounts of pure protein that remains stable in a buffered solution is a strict requirement for crystallisation trials and therefore removal of flexible disordered regions that impeded the crystallisation process could be targeted for removal (Deller et al., 2016; Pantazatos et al., 2004).

A miniaturised expression experiment was designed to compare the apparent yield and solubility of recombinantly expressed C-terminally truncated and full length RepC proteins in parallel using a range of IPTG concentrations for induction. Miniaturisation facilitated rapid screening of multiple expression clones whilst significantly reducing the amounts of chemical reagents used for expression trials. Miniaturisation and parallelisation have gained considerable attention in exploratory recombinant protein expression. Application of these principles is particularly useful for the expression of novel or difficult proteins. A broad range of expression parameters can be tested to identify promising conditions which could be applied to large scale downstream expression experiments (Hunt, 2005). In the current miniaturised expression trial, it was evident that truncation of His-RepC significantly improved cell pellet yield and protein expression levels. Both proteins were predicted to be unstable and were expressed in the soluble fraction. This was not expected since His-RepC was expressed in the soluble fraction in previous large scale expression trials. It is possible that some extrinsic factors including, but not limited to, heating during sonication of smaller sample volumes may have negatively impacted solubility (Kramer et al., 2012). Large scale expression of His-RepC_14 is recommended to ascertain if expression

levels are increased accordingly when scaling up expression culture volumesWhilst expression levels of MBP-RepC and MBP-RepC_14 did not appear to differ, the RepC_14 protein was predicted to be stable after tag cleavage. Although additional downstream purification and activity assays are necessary to determine if truncation of RepC will affect its activity in vitro it is vital to consider that the predicted crystallisability of RepC remains low despite efforts to improve yield and stability. Although X-ray crystallography is considered the gold standard of protein structure characterisation, it is not without its limitations particularly for highly dynamic proteins such as RepC (McPherson and Gavira, 2014). It has been established that for begomovirus Rep to display helicase activity in vitro it must display at least a dodecameric conformation (Choudhury et al., 2006; Clérot and Bernardi, 2006). As a result, the biologically relevant structure of this protein may display a degree of complexity exceeding that of other characterised viral helicases. The recent demonstration of the complex conformational shifts in the Rep protein of AAV2 using cryo-EM (Santosh et al., 2020) further emphasise that attempts to crystallise an extremely dynamic protein such as begomovirus Rep may prove to be a challenging endeavour. Recent advances in cryo-EM technology which have resulted in significantly enhanced structure resolution (Assaiya et al., 2021; Kühlbrandt, 2014; Rivera-Calzada and Carroni, 2019) may provide the perfect opportunity to structurally characterise this intriguing protein. This study highlights only a few of the expression parameters that require ample optimisation steps to achieve consistent results. Despite these challenges, the unique characteristics of begomovirus Rep make this protein a worthwhile target for structural studies of viral helicases.

The present study was the first to identify the functional motifs of the Rep protein of ToCSV V30 and analyse the predicted structure using in silico prediction methods and homology modelling. The V30 Rep protein was shown to contain the conserved functional motifs characteristic of the HUH endonuclease and ATPase/SF3 helicase domains of related viral Rep proteins. The arginine finger motif was absent in V30 Rep which is in agreement with other structural studies of geminivirus Rep proteins that were shown to lack this motif. This study also demonstrated that the V30 C1 putative 5’ extension region was transcribed in vivo which suggests that an N-terminally extended Rep protein may be produced but further studies are needed to confirm the existence of multiple forms of Rep and the biological relevance thereof. It was

established that Rep aa 346 – 359 are not required for viral infectivity in vivo although it is hypothesised that the delay in disease symptom development was due to impaired function of Rep lacking this region. Additional experiments are required to assess the in vitro ATPase and helicase activity of C-terminal deletion mutants. Further deletion of Rep aa 347 – 338 resulted in abolished infectivity which indicates that these residues are indispensable to infection although the specific activity of Rep that was disrupted by this truncation remains to be determined. This study was also the first to report the successful expression of V30 RepC in the soluble fraction using pCold expression vectors in *E. coli*. Tag cleavage was successfully conducted, and purification experiments showed some success although further optimisation is necessary to improve purity. Through small scale expression trials of truncated RepC proteins, a promising modification to increase protein yield and predicted stability was identified. Targeted removal of disordered regions may greatly improve the chances of obtaining larger quantities of purified Rep for future structural studies.

Chapter 4: Concluding remarks

4.1 Brief rationale and primary aims of the study

Cultivated plant crops such as tomato are important to the commercial and subsistence farming economies of Africa. The continued development of agriculture on the continent is threatened by virus diseases spread by hemipteran insect pests. The reported diversity of whitefly-transmitted geminiviruses present in both cultivated and wild plant species in Africa is a serious cause for concern and evidence of their continued evolution warrants increased study of these viruses (Lefeuvre et al., 2007a; Leke et al., 2015; Mivedor et al., 2017; Rey et al., 2012). Surveys of tomato growing regions of Southern Africa have identified ToCSV as an important pathogen in this region (Esterhuizen, 2012; Moodley et al., 2019; Sande et al., 2021). Two ToCSV variant groups have been identified in SA (ToCSV-I and ToCSV-II) and are differentiated by the presence of a recombinant fragment in the IR and V2 regions of ToCSV-II group isolates and by the differing severity of disease symptoms induced in tomato (Esterhuizen, 2012). To investigate the differing pathologies further in the present study, infectivity assays were conducted for the first time in the experimental host plant *Nicotiana benthamiana* using swap mutant chimeras. The ToCSV Rep was included in this study since it is a multifunctional protein that is crucial for successful replication in the host (Hanley-Bowdoin et al., 2013; Rizvi et al., 2015). No previous studies of the ToCSV Rep protein are available and this pioneering study of African begomovirus Rep proteins was conducted with the hope of inspiring greater study of Rep in Africa.

The primary aims of the current work were to explore and compare the coding capacity of ToCSV and related African begomoviruses in silico, examine the role that sequence differences exert in the pathogenicity of two variants in *N. benthamiana*, and investigate the structure and function of the ToCSV Rep protein using various techniques.

4.2 Major study findings

Based on the observations and data obtained in the present study, the major findings are summarised as follows:

1. Five novel putative ORFs encoding proteins ≥40 aa were identified in V30, V22 and related mono- and bipartite African begomoviruses. The C5/AC5 and C6/AC6 ORFs were the most commonly identified putative ORFs. The C7 ORF was found in monopartite begomoviruses only.
2. The symptom phenotypes induced by V30 and V22 in *N. benthamiana* were distinguished in two ways. Firstly, V22 induced moderate to severe ULR which was largely absent in plants infected with V30. Secondly, a recovery phenotype in V22-infected plants was observed as a decrease in ULR and SV symptom severity between 24 and 36 dpi, whereas SV symptom severity in V30-infected plants did not decrease over this period.
3. A peak in viral titre was recorded at 20 dpi in plants infected with either V30 or V22, and viral load was significantly higher in V30-infected plants than V22-infected plants at peak and late infection timepoints.
4. RNA transcripts derived from both the vs and cs strands were identified in V30-infected *N. benthamiana* using RT-PCR with strand-specific primers. These transcripts spanned multiple ORFs including all putative C5, C6, C7, V3 and V4 ORFs. In addition, IR-derived vs and cs transcripts crossing the Ori and transcripts derived from putative V2 and C1 5' ORF extensions were also positively identified.
5. Agroinoculation of *N. benthamiana* with V30 IR swap mutant chimeras demonstrated that sequence differences present in the 3' IR are responsible for the differing symptom phenotypes induced by V30 and V22. The 3' IR sequence between the TATA box promoter and the V2 start codon is particularly important. The presence of the V22 sequence spanning this short region in V30 was sufficient to result in significantly increased ULR symptom development.
6. Through agroinoculation of *N. benthamiana* with V30 and V22 V2 ORF swap mutant chimeras, it was determined that sequence differences present in the canonical V2 ORF are responsible for the modulating both viral titre, disease symptom severity and V22-specific disease recovery phenotype.

7. Using multiple aa sequence sequence alignment, a novel putative domain termed the NLHr which was predicted to contain a transmembrane domain was identified in the C5/AC5 proteins of African monopartite and bipartite begomoviruses. A novel putative C6/AC6 domain designated the C6cd was also identified.
8. Agroinoculation of *N. benthamiana* with a C6 ORF deletion mutant revealed that the 35 C-terminal aa of V30 C6 are not required for successful infection although this region may contribute to influencing SV symptom severity through an unknown mechanism. The positive identification of an RNA transcript derived from the C6 coding region is evidence that this putative ORF is transcribed in infected *N. benthamiana*.
9. The positions of important HUH endonuclease motifs (motif I, II, III) and ATPase/SF3 helicase motifs (WA, WB, B' motif, C motif) were mapped for the Rep proteins of ToCSV and other African begomoviruses. Additionally, the GRS and IRD sequence motifs were also identified. The arginine finger motif was absent.
10. Homology modelling of ToCSV Rep showed that the N-terminal endonuclease domain was predicted to contain an antiparallel β-sheet (five β-strands) whilst the C-terminal ATPase/helicase domain contained a parallel β-sheet (four β-strands), and the arrangement of functional motifs was similar to that present in the solved structures of other DNA virus Rep proteins. The ATPase/helicase domain was modelled as a homohexamer.
11. Agroinoculation of *N. benthamiana* with V30 C1 deletion mutants revealed that removal of 14 C-terminal residues (aa 346 – 359) did not abolish infectivity but resulted in significantly reduced viral titre and delayed symptom development. C1 deletion mutants encoding a Rep protein lacking 22 C-terminal residues (aa 338 – 359) were unable to infect *N. benthamiana*.
12. ToCSV V30 RepC (aa 122 – 359) was successfully expressed as a recombinant fusion in the soluble fraction using pCold I, pCold TF, and pMAL-c2X expression vectors in *Escherichia coli*. Tag cleavage and partial purification was achieved. Miniaturised expression trials of RepC truncated by 14 C-terminal residues resulted in an improved apparent protein yield.

4.3 Implications

A recent study of small putative ORFs present in TYLCV (Gong et al., 2021) suggests that small non-canonical proteins have diverse subcellular localisations and may contribute to modulating virus infection. This important research highlights that the coding capacity of geminiviruses has been underappreciated and the putative ORFs identified in the current study deserve serious attention. Since the C5/AC5, C6/AC6, and C7 proteins identified in this study aligned with putative proteins identified by Gong et al. (2021) it is proposed that these putative ORFs are unlikely to be spurious artefacts. The role of C5/AC5 as a suppressor of RNA silencing has been established (Li et al., 2015; Li et al., 2021b) although the precise function of the AC5-2 domain has not been determined. The predicted TMH in the NLHr region of C5/AC5 may further modulate C5/AC5 activity by altering subcellular localisation. This study was the first to identify a putative domain in the C6/AC6 protein and regions of sequence conservation suggest that this protein may serve a key function in diverse begomovirus isolates. It is proposed that the presence of internal methionine codons in the C5/AC5 and C6/AC6 ORFs could result in the expression of multiple forms of these proteins. Should this be the case, the variety of potential functions attributed to these proteins will undoubtedly increase. Since the C7 ORF was found in monopartite begomoviruses only, this suggests that it may be involved in activity related to viral movement that is associated with BV1 and/or BC1 in bipartite begomoviruses. This is further supported by the observation that the C7 homologue of TYLCV was predicted to contain a transmembrane domain and localised to the ER (Gong et al., 2021). Another component related to monopartite begomovirus movement is therefore suggested to exist.

The current study has also determined that there is an urgent need to develop a system of nomenclature for putative begomovirus ORFs especially since not all begomovirus isolates examined in this study contained both C5/AC5 and C6/AC6 ORFs and some encoded multiple C5/AC5 protein segments. These phenomena have resulted in inconsistent putative ORF designations in previous studies (mentioned in Chapter 2). Continued inconsistency in assigning names to putative ORFs present in the V1/V2 coding region will inevitably result in further confusion. It is proposed that the cs ORF in frame with V1/AV1 is designated C5/AC5 and the cs ORF in frame with V2/AV2 is designated C6/AC6. Additional sequence analysis is necessary to confirm

subsequent ORF designations. Regardless of their nomenclature, the presence of small putative ORFs signify that several interaction pathways remain to be described and the competitive advantage that these ORFs may serve in mixed infection adds another layer of complexity to the puzzle. Perhaps these ORFs are remnants from the evolution of these viruses in wild non-cultivated hosts? Until their role in infection is better understood, one can only speculate as to how begomoviruses have evolved to contain these diverse and seemingly inconspicuous ORFs.

The importance of the recombinant fragment present in V22 was demonstrated in the present study. Sequence differences in this region were responsible for differentiating symptom phenotype, severity, and recovery in infected *N. benthamiana*. The recent discovery of a conserved geminivirus TACE (Cantú-Iris et al., 2019) suggests that this novel element requires consideration when investigating the expression of vs ORFs. The observed changes in disease symptom phenotype induced by IR swap mutants were linked to sequence differences present in a short IR region containing the TACE. Since the V30 TACE does not conform to the arrangement described for OW begomoviruses (Cantú-Iris et al., 2019) it provides an opportunity to study the role of this element in disease symptom phenotype determination. The results of the present study and that conducted by Brigden (2022) indicate that sequence differences in the N-terminal region of the V2 protein are responsible for modulating disease recovery in *N. benthamiana* infected with V22. Residues present in this region that were previously determined to be important in the RNA silencing suppression activity of V2 (Li et al., 2021a) were conserved between the two ToCSV variant V2 proteins. The sequence difference at V2 aa position 27 was shown to be critical to altered disease severity and recovery (Brigden, 2022). Although no studies investigating this residue have been previously published, Basu et al. (2021) hypothesised that sequence differences outside of known conserved motifs may exert a significant influence on the activity of closely related V2 proteins (Basu et al., 2021). This strongly suggests that more studies are necessary to determine the role that other residues present in this region play in pathways linked to RNA silencing suppression. The atypical serine present in the putative PKC phosphorylation motif of V30 which was rarely found in African begomovirus V2 proteins serves as an exception to the previously described PKC consensus sequence (Chowda-Reddy et al., 2008; Mubin et al., 2010; Saeed et al., 2018). This finding implies that the threonine of the PKC motif is not a strict

requirement for V2 function and substitution of this residue with serine in the PKC motif of V22 V2 resulted in delayed symptom development but no change in viral titre or recovery levels (Brigden, 2022). Additional studies are needed to determine how the function of this motif could be regulated by deviations from the consensus sequence. The positive detection of long RNA transcripts in ToCSV-infected *N. benthamiana* is evidence of possible readthrough transcription which is an important event in the production of siRNAs which are integral to RNA silencing pathways in plants (Pooggin, 2013).

Although the Rep of ToCSV was modelled as a hexamer, it is important to consider that begomovirus Rep proteins that previously displayed helicase activity in vitro adopted a range of higher order oligomeric conformations (dodecamer to 24-mer) (Choudhury et al., 2006; Clérot and Bernardi, 2006). Although homology modelling may provide important information regarding the arrangement of conserved functional motif residues, it must be acknowledged that geminivirus Rep helicase domain might adopt a unique fold not present in other solved SF# helicase structures. This is relevant especially since geminivirus Rep lacks the arginine finger motif which is implicated in the coordination of adjacent *trans*-monomers in the oligomeric structures of other SF3 helicases (Enemark and Joshua-Tor, 2006; Gai et al., 2004). This highlights the need to solve the structure of geminivirus Rep to obtain a more biologically relevant model of how these unique SF3 helicases function. The investigation of the C-terminal domain in this study has two important implications. Firstly, the results of in vivo experiments suggest that the predicted disordered region spanning 14 C-terminal residues of Rep is not a strict requirement for enzymatic activity required for infection. Secondly, targeted removal of this disordered region may significantly improve yield of recombinantly expressed protein. The boundaries of the C-terminal domain may need to be redefined. The N-terminal part of this domain is predicted to form part of the parallel β-sheet of the helicase domain whilst the C-terminal end is evidently distinct from this catalytic core region. It is possible that the C-terminal end may function as a flexible C-terminal tail especially since it contains a series of acidic residues. The C-tT is present in other SF3 helicases and has been shown to function in stabilising the oligomeric helicase structure (Javed et al., 2021). How this region could be involved in stabilising a higher order oligomeric structure of geminivirus Rep remains to be determined.

4.4 Challenges and Limitations

One of the significant limitations encountered in this study was the inability to mutate residues of interest present in the putative ORFs due to considerable sequence overlap present in the V1/V2 coding region. A comprehensive assessment of the C6 ORF revealed that mutagenesis of the ORF region encoding conserved residues of interest in the C6cd was not possible without introducing inadvertent mutations to overlapping ORFs. This issue was also encountered when investigating mutation of the ToCSV C5 ORF region encoding the AC5-2 domain. Furthermore, the presence of multiple methionine codons in the C6 ORF prevented the generation of a full C6 ORF deletion mutant. The presence of significant ORF overlap was a confounding factor in interpretation of the results of pathogenicity trials involving the V30 and V22 V2 ORF swap mutants. This issue was partly addressed by Brigden (2022) via generation of V2 substitution mutants without inadvertent mutation to overlapping ORFs. It was also determined by Brigden (2022) however, that certain sequence differences present in the V2 could not be targeted for substitution because of introduction of inadvertent mutations to the putative C6 ORF which overlaps with the 5’ region of V2.

Although quantitative assessment of vs ORF transcript levels would have been a beneficial addition to this study, reference gene validation has not been previously conducted for *N. benthamiana* infected with ToCSV. Since a range of reference genes should be validated prior to quantitative RT-PCR (Bokhale et al., 2020; Liu et al., 2012) it was not possible to conduct this experiment due to funding and time constraints. Deep sequencing of siRNAs in the present study could also not be investigated due to funding constraints. Conducting infectivity trials in tomato using the mutant chimeras was not possible due to the lack of adequate growth facility space for tomato which requires significantly more space than *N. benthamiana*. It is however recognised that future studies should be conducted in the natural host of this virus to determine the impact of sequence differences on pathology and infectivity in tomato.

It is acknowledged that the homology modelling of the helicase domain of ToCSV Rep was impeded by the lack of homologues in the PDB. The templates identified shared low sequence identity with ToCSV Rep (<20 %). This was especially true for the regions outside of the conserved functional motifs which were highly divergent in both length and sequence identity. In the case of Phyre2-generated homology model, some

regions of the C-terminal and helicase domains were modelled using ab initio methods which are not particularly reliable. This further emphasises the need to solve the structure of geminivirus Rep helicase domain. Experiments aimed at recombinant expression of ToCSV RepC were met with challenges at various stages. Inconsistent levels of protein expression were obtained particularly with the pCold I expression vector. The inherent toxicity of RepC gene and/or protein prevented the use of *E. coli* expression strains that are used to express high levels of target protein. Access to liquid chromatography systems necessary for purification was interrupted throughout the study due to several unforeseen circumstances. Reliance on gravity flow purification was detrimental to the present study due to lower processing volumes and extended purification times. The lack of an automated purification system impeded the range of optimisation allowed in the time frames allocated to this component of the study.

4.5 Questions and future work

Some important questions that have arisen as a result of this study include:

What role do the identified putative ORFs play in infection and why are some rarer than others?

Why are the lengths of C5/AC5 and C6/AC6 ORFs highly variable whilst the lengths of C7, V3/AV3 and V4/AV4 are less so?

Are these putative ORFs found in the DNA-A of other geminivirus genera and if so, how prevalent are they?

What is the specific function of the conserved NLHr identified in C5/AC5 proteins and the C6cd identified in C6/AC6 proteins?

Is the protein encoded by the C7 ORF present in monopartite begomoviruses involved in virus movement? Does it function as a homologue of MP encoded by DNA-B of bipartite begomoviruses?

What is the relevance of the identified vs and cs strand-derived RNA transcripts in infection?

What is the specific mechanism which is responsible for the different symptom phenotypes induced by the IR swap mutants?

What is the role of the atypical TACE present in the 3' IR of V30? Why is ToCSV the only African begomovirus identified in this study with such a unique TACE sequence?

What function do specific aa sequence differences present in the N-terminal region of V30 and V22 V2 have that modulate disease severity and recovery?

How does the atypical serine present in the PKC motif of V30 V2 modify its presumed function and/or activity in RNA silencing suppression?

What is the role of the putative V2 and C1 ORF 5' extensions and are they translated?

What sequence motif carries out the function of the missing arginine finger motif in the geminivirus helicase domain?

What is the specific role of the disordered region of the C-terminal domain in ToCSV pathogenicity in *N. benthamiana*?

The present study has provided several avenues for future work and various experiments are recommended to expand on the results obtained herein. In silico searches of putative ORFs should be extended to include other OW and NW begomoviruses to increase the number of genomes investigated. By including more diverse geminiviruses, the prevalence and diversity of these ORFs can be better understood. To identify the ends of the putative ORF transcripts, 5' and 3' RACE assays are recommended (Akbar et al., 2012). This will allow precise mapping of the transcripts identified in this study. In conjunction with promoter mapping techniques (Khan et al., 2015; Shivaprasad et al., 2005), the likelihood that these putative ORFs are translated could be evaluated. Ribosome profiling could prove extremely useful in revealing the various translation events that take place in begomovirus-infected plants especially in the absence of information regarding putative ORF promoter elements (Brar and Weissman, 2015; Stern-Ginossar, 2015).

Deep sequencing would be useful to compare the vsRNA profiles of the two variants in infected *N. benthamiana* at timepoints of interest such as peak infection (20 dpi), or late infection (36 dpi) when recovery is evident in V22-infected plants. By comparing these profiles, the significance of "hotspot" regions (which contain sequence

differences) can be assessed (Miozzi et al., 2013; Piedra-Aguilera et al., 2019). To quantify viral gene expression, quantitative RT-PCR is suggested. This is particularly relevant in the case of plants infected with V30 IR swap mutants which are hypothesised to have differential late vs gene promoter activity influenced by sequence differences present in the vicinity of rightward promoters. Prior to quantitative RT-PCR assays, validation of reference genes is required to ensure normalisation of data. Although this has been previously done for tomato infected with ToCSV (Bokhale et al., 2020) reference gene validation remains to be done for *N. benthamiana* infected with ToCSV.

Due to considerable sequence overlap of putative cs ORFs in the V1/V2 coding region, site-directed mutagenesis of these ORFs in the V30 and V22 infectious clones may not be possible without introducing inadvertent mutations to overlapping ORFs. It will be worthwhile to investigate the functional characteristics of each putative ORF. The subcellular localisation of each protein can be determined by transient expression as green fluorescent protein (GFP)-fusions (Gong et al., 2021; Zhan et al., 2018). This may provide insight into the potential function of each protein. To establish whether the putative proteins act as pathogenicity determinants in *N. benthamiana*, PVX-vector based expression could be conducted (Sharma and Ikegami, 2010; Zhan et al., 2018). The ability of the putative ORFs to suppress PTGS in the GFP-transgenic *N. benthamiana* 16c line could be assessed using the methods described by Yang et al. (2018) and Zhan et al. (2018). These techniques can be extended to evaluate V2 and C1 ORFs containing putative 5' extensions and putative ORF mutants to investigate the role of selected residues in these activities. It is imperative that additional infectivity trials with ToCSV wild type infectious clone and mutant chimeras are conducted in different cultivars of tomato. Viral load determination using qPCR will allow for quantitative assessment of viral load in tomato infected with V30 and V22 which remains to be determined. Of particular interest is whether the IR swap mutants induce altered symptom phenotypes in tomato.

Future experiments aimed at recombinant expression and structural characterisation of ToCSV RepC require significant optimisation. Alternative expression systems using different tags and tag cleavage sites should be considered (Kimple et al., 2013; Lichty et al., 2005; Zhao et al., 2013). Tag cleavage should be conducted at low temperatures to limit protein aggregation and precipitation. Expression of RepC in alternative *E. coli*

strains that are more suited for expression of toxic proteins such as C41(DE3) or C43(DE3) may improve yield and/or solubility (Dumon-Seignovert et al., 2004; Miroux and Walker, 1996). The results obtained in this study suggest that removal of disordered regions in the C-terminal domain may significantly improve yield and stability of the recombinantly expressed protein. Enzymatic activity assays should be conducted to confirm that truncated proteins retain functionality in vitro. Since RepC is predicted to be unlikely to crystallise and forms higher order oligomers, it is recommended that cryogenic electron microscopy (cryo-EM) should be seriously considered as a technique to solve the structure of the Rep helicase domain. This method has recently proved particularly useful to investigate multiple protein domains, flexible regions, and conformational variability of other DNA virus Rep proteins (Javed et al., 2021; Santosh et al., 2020; Tarasova et al., 2021). It is hoped that the present study will serve to provide several avenues for future research into begomovirus infection of tomato.

References

Abbas, Q., Amin, I., Mansoor, S., Shafiq, M., Wassenegger, M. and Briddon, R.W., 2019. The Rep proteins encoded by alphasatellites restore expression of a transcriptionally silenced green fluorescent protein transgene in Nicotiana benthamiana. VirusDisease, 30(1), pp.101-105. https://doi.org/10.1007/s13337-017-0413-5

Abbate, E.A., Berger, J.M. and Botchan, M.R., 2004. The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. Genes Dev., 18(16), pp.1981-1996. https://doi.org/10.1101/gad.1220104

Ahmed, N., Mahmood, M.A., Amin, I. and Mansoor, S., 2021. Geminiviruses also encode small proteins with specific functions. Trends Microbiol., 29(12), pp.1052-1054. https://doi.org/10.1016/j.tim.2021.09.008

Akbar, F., Briddon, R.W., Vazquez, F. and Saeed, M., 2012. Transcript mapping of Cotton leaf curl Burewala virus and its cognate betasatellite, Cotton leaf curl Multan betasatellite. Virol. J., 9, 249 . https://doi.org/10.1186/1743-422X-9-249

Akbergenov, R., Si-Ammour, A., Blevins, T., Amin, I., Kutter, C., Vanderschuren, H., Zhang, P., Gruissem, W., Meins Jr, F., Hohn, T. and Pooggin, M.M., 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res., 34(2), pp.462-471. https://doi.org/10.1093/nar/gkj447

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol., 215(3), pp.403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amin, I., Patil, B.L., Briddon, R.W., Mansoor, S. and Fauquet, C.M., 2011. Comparison of phenotypes produced in response to transient expression of genes encoded by four distinct begomoviruses in *Nicotiana benthamiana* and their correlation with the levels of developmental miRNAs. Virol. J., 8, 238. https://doi.org/10.1186/1743-422x-8-238

Angius, F., Ilioaia, O., Amrani, A., Suisse, A., Rosset, L., Legrand, A., Abou-Hamdan, A., Uzan, M., Zito, F. and Miroux, B., 2018. A novel regulation mechanism of the T7 RNA polymerase based expression system improves overproduction and folding of membrane proteins. Sci. Rep., 8, 8572. https://doi.org/10.1038/s41598-018-26668-y

Ansar, M., Agnihotri, A.K., Akram, M. and Bhagat, A.P., 2019. First report of Tomato leaf curl Joydebpur virus infecting French bean (Phaseolus vulgaris L.). J. Gen. Plant Pathol., 85(6), pp.444-448. https://doi.org/10.1007/s10327-019-00867-5

Aregger, M., Borah, B.K., Seguin, J., Rajeswaran, R., Gubaeva, E.G., Zvereva, A.S., Windels, D., Vazquez, F., Blevins, T., Farinelli, L. and Pooggin, M.M., 2012. Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog. 8, e1002941 https://doi.org/10.1371/journal.ppat.1002941

Argüello-Astorga, G.R., Guevara-Gonzalez, R.G., Herrera-Estrella, L.R. and Rivera-Bustamante, R.F., 1994. Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. Virology, 203(1), pp.90-100. https://doi.org/10.1006/viro.1994.1458

Argüello-Astorga, G.R. and Ruiz-Medrano, R., 2001. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-base pairs by a comparative approach. Arch. Virol., 146(8), pp.1465-1485. https://doi.org/10.1007/s007050170072

Argüello-Astorga, G., Lopez-Ochoa, L., Kong, L.J., Orozco, B.M., Settlage, S.B. and Hanley-Bowdoin, L., 2004. A novel motif in geminivirus replication proteins interacts with the plant retinoblastoma-related protein. J. Virol., 78(9), pp.4817-4826. https://doi.org/10.1128/JVI.78.9.4817-4826.2004

Armenteros, J.J.A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A. and Nielsen, H., 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance, 2, e201900429 https://doi.org/10.26508/lsa.201900429

Ashfaq, M.A., Dinesh Kumar, V., Soma Sekhar Reddy, P., Anil Kumar, C., Sai Kumar, K., Narasimha Rao, N., Tarakeswari, M. and Sujatha, M., 2020. Post-transcriptional gene silencing: Basic concepts and applications. J. Biosci., 45(1), pp.1-10. https://doi.org/10.1007/s12038-020-00098-3

Ashraf, M.A., Shahid, A.A., Rao, A.Q., Bajwa, K.S. and Husnain, T., 2014. Functional characterization of a bidirectional plant promoter from cotton leaf curl Burewala virus using an Agrobacterium-mediated transient assay. Viruses, 6(1), pp.223-242. https://doi.org/10.3390/v6010223

Assaiya, A., Burada, A.P., Dhingra, S. and Kumar, J., 2021. An overview of the recent advances in cryo-electron microscopy for life sciences. Emerg. Top. Life Sci., 5(1), pp.151-168. https://doi.org/10.1042/ETLS20200295

Bañuelos-Hernández, B., Mauricio-Castillo, J.A., Cardenas-Conejo, Y., Guevara-González, R.G. and Arguello-Astorga, G.R., 2012. A new strain of tomato severe leaf curl virus and a unique variant of tomato yellow leaf curl virus from Mexico. Arch. Virol., 157(9), pp.1835-1841. https://doi.org/10.1007/s00705-012-1358-z

Bar-Ziv, A., Levy, Y., Hak, H., Mett, A., Belausov, E., Citovsky, V. and Gafni, Y., 2012. The Tomato yellow leaf curl virus (TYLCV) V2 protein interacts with the host papain-like cysteine protease CYP1. Plant Signal. Behav., 7(8), pp.983-989. https://dx.doi.org/10.4161%2Fpsb.20935

Bar-Ziv, A., Levy, Y., Citovsky, V. and Gafni, Y., 2015. The Tomato yellow leaf curl virus (TYLCV) V2 protein inhibits enzymatic activity of the host papain-like cysteine protease CYP1. Biochem. Biophys. Res. Commun., 460(3), pp.525-529. https://doi.org/10.1016/j.bbrc.2015.03.063

Basu, S., Kumar Kushwaha, N., Kumar Singh, A., Pankaj Sahu, P., Vinoth Kumar, R. and Chakraborty, S., 2018. Dynamics of a geminivirus-encoded pre-coat protein and host RNA-dependent RNA polymerase 1 in regulating symptom recovery in tobacco. J. Exp. Bot., 69(8), pp.2085-2102. https://doi.org/10.1093/jxb/ery043

Basu, S., Singh, A.K., Singh, D., Sahu, S.K. and Chakraborty, S., 2021. Role of viral suppressors governing asymmetric synergism between tomato-infecting begomoviruses. Appl. Microbiol. and Biotechnol., 105(3), pp.1107-1121. https://doi.org/10.1007/s00253-020-11070-4

Bernardo, P., Golden, M., Akram, M., Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A.G., Peterschmitt, M., Martin, D.P. and Roumagnac, P., 2013. Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. Virus Res., 177(1), pp.35-45. https://doi.org/10.1016/j.virusres.2013.07.006

Berrie, L.C., Palmer, K.E., Rybicki, E.P. and Rey, M.E.C., 1998. Molecular characterisation of a distinct South African cassava infecting geminivirus. Arch. Virol., 143(11), pp.2253-2260. https://doi.org/10.1007/s007050050457

Berrie, L.C., Rybicki, E.P. and Rey, M.E.C., 2001. Complete nucleotide sequence and host range of South African cassava mosaic virus: further evidence for recombination amongst begomoviruses. J. Gen. Virol., 82(1), pp.53-58. https://doi.org/10.1099/0022-1317-82-1-53

Bigarré, L., Chazly, M., Salah, M., Ibrahim, M., Padidam, M., Nicole, M., Peterschmitt, M., Fauquet, C. and Thouvenel, J.C., 2001. Characterization of a new begomovirus from Egypt infecting hollyhock (Althea rosea). Eur. J. Plant Pathol., 107(7), pp.701-711. https://doi.org/10.1023/A:1011967232319

Bisaro, D.M., 2006. Silencing suppression by geminivirus proteins. Virology, 344(1), pp.158-168. https://doi.org/10.1016/j.virol.2005.09.041

Blancard, D., Laterrot, H., Marchoux, G. and Candresse, T., 2012. The tomato plant and its culture. In: Fletcher, J. (ed.) *Tomato Diseases: Identification, Biology and Control,* 2nd ed, Manson Publishing, London, pp.17-35.

Bokhale, M., Mwaba, I. and Allie, F., 2020. Real-time PCR data for reference candidate gene selection in tomato infected with Tomato curly stunt virus. Data Brief, 31, p.105750. https://doi.org/10.1016/j.dib.2020.105750

Borah, B.K., Zarreen, F., Baruah, G. and Dasgupta, I., 2016. Insights into the control of geminiviral promoters. Virology, 495, pp.101-111. https://doi.org/10.1016/j.virol.2016.04.033

Boulton, M.I., King, D.I., Donson, J. and Davies, J.W., 1991. Point substitutions in a promoter-like region and the V1 gene affect the host range and symptoms of maize streak virus. Virology, 183(1), pp.114-121. https://doi.org/10.1016/0042-6822(91)90124-t

Brar, G.A. and Weissman, J.S., 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nature reviews Mol. Cell Biol., 16(11), pp.651-664. https://doi.org/10.1038/nrm4069

Brázda, V., Laister, R.C., Jagelská, E.B. and Arrowsmith, C., 2011. Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol. Biol., 12, 33. https://doi.org/10.1186/1471-2199-12-33

Briddon, R.W., Martin, D.P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olivé, E., Moriones, E., Lett, J.M., Zerbini, F.M. and Varsani, A., 2018. Alphasatellitidae: a new family with two subfamilies for the classification of geminivirus-and nanovirus-associated alphasatellites. Arch. Virol., 163(9), pp.2587-2600. https://doi.org/10.1007/s00705-018-3854-2

Briddon, R. W., Ghabrial, S., Lin, N.-S., Palukaitis, P., Scholthof, K. B. G., and Vetten, H. J. 2012. Satellites and other virus-dependent nucleic acids. In: A. M. Q. King, E. Lefkowitz, M. J. Adams, and E. B. Carstens (eds) *Virus Taxonomy - Ninth Report of the International Committee on Taxonomy of Viruses*, Associated Press, New York, pp.1209–1219.

Brigden, A., 2022. MSc thesis. Pathogenicity determinants of severe and mild variants of Tomato curly stunt virus (ToCSV). University of the Witwatersrand (South Africa).

Brown, N.P., Leroy, C. and Sander, C., 1998. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics, 14(4), pp.380-381. https://doi.org/10.1093/bioinformatics/14.4.380

Brown, J.K., Ostrow, K.M., Idris, A.M. and Stenger, D.C., 2000. Chino del tomate virus: relationships to other begomoviruses and identification of A-component variants that affect symptom expression. Phytopathology, 90(5), pp.546-552. https://doi.org/10.1094/PHYTO.2000.90.5.546

Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C., Fiallo-Olivé, E., Briddon, R.W., Hernández-Zepeda, C., Idris, A. and Malathi, V.G., 2015. Revision of Begomovirus taxonomy based on pairwise sequence comparisons. Arch. Virol., 160(6), pp.1593-1619. https://doi.org/10.1007/s00705-015-2398-y

Buchan, D.W. and Jones, D.T., 2019. The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res., 47(W1), pp.W402-W407. https://doi.org/10.1093/nar/gkz297

Caicedo, A. and Peralta, I., 2013. Basic information about tomatoes and the tomato group. In: Liedl, B. E., Labate J. A., Stommel, J. R., Slade, A., and Kole, C. (eds.) *Genetics, Genomics and Breeding of Tomato*, CRC Press, Boca Raton, pp.1-36.

Campos-Olivas, R., Louis, J.M., Clérot, D., Gronenborn, B. and Gronenborn, A.M., 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. Proc. Natl. Acad. Sci., 99(16), pp.10310-10315. https://doi.org/10.1073/pnas.152342699

Cantú-Iris, M., Pastor-Palacios, G., Mauricio-Castillo, J.A., Bañuelos-Hernández, B., Avalos-Calleros, J.A., Juárez-Reyes, A., Rivera-Bustamante, R. and Argüello-Astorga, G.R., 2019. Analysis of a new begomovirus unveils a composite element conserved in the CP gene promoters of several Geminiviridae genera: Clues to comprehend the complex regulation of late genes. PloS one, 14(1), p.e0210485. https://doi.org/10.1371/journal.pone.0210485

Castillo, A.G., Collinet, D., Deret, S., Kashoggi, A. and Bejarano, E.R., 2003. Dual interaction of plant PCNA with geminivirus replication accessory protein (Ren) and viral replication protein (Rep). Virology, 312(2), pp.381-394. https://doi.org/10.1016/S0042-6822(03)00234-4

Castillo-González, C., Liu, X., Huang, C., Zhao, C., Ma, Z., Hu, T., Sun, F., Zhou, Y., Zhou, X., Wang, X.J. and Zhang, X., 2015. Geminivirus-encoded TrAP suppressor inhibits the histone methyltransferase SUVH4/KYP to counter host defense. Elife, 4, p.e06671. https://doi.org/10.7554/eLife.06671

Chaban, Y., Stead, J.A., Ryzhenkova, K., Whelan, F., Lamber, E.P., Antson, A., Sanders, C.M. and Orlova, E.V., 2015. Structural basis for DNA strand separation by a hexameric replicative helicase. Nucleic Acids Res., 43(17), pp.8551-8563. https://doi.org/10.1093/nar/gkv778

Chakraborty, S., Vanitharani, R., Chattopadhyay, B. and Fauquet, C.M., 2008. Supervirulent pseudorecombination and asymmetric synergism between genomic components of two distinct species of begomovirus associated with severe tomato leaf curl disease in India. J. Gen. Virol., 89(3), pp.818-828. https://doi.org/10.1099/vir.0.82873-0

Chatterji, A., Padidam, M., Beachy, R.N. and Fauquet, C.M., 1999. Identification of replication specificity determinants in two strains of tomato leaf curl virus from New Delhi. J. Virol., 73(7), pp.5481-5489. https://doi.org/10.1128/JVI.73.7.5481-5489.1999

Chellappan, P., Vanitharani, R. and Fauquet, C.M., 2004. Short interfering RNA accumulation correlates with host recovery in DNA virus-infected hosts, and gene silencing targets specific viral sequences. J. Virol., 78(14), pp.7465-7477. https://doi.org/10.1128/jvi.78.14.7465-7477.2004

Chellappan, P., Vanitharani, R. and Fauquet, C.M., 2005. MicroRNA-binding viral protein interferes with Arabidopsis development. Proc. Natl. Acad. Sci. USA, 102(29), pp.10381-10386. https://doi.org/10.1073/pnas.0504439102

Choudhury, N.R., Malik, P.S., Singh, D.K., Islam, M.N., Kaliappan, K. and Mukherjee, S.K., 2006. The oligomeric Rep protein of Mungbean yellow mosaic India virus (MYMIV) is a likely replicative helicase. Nucleic Acids Res., 34(21), pp.6362-6377. https://doi.org/10.1093/nar/gkl903

Chowda-Reddy, R.V., Achenjang, F., Felton, C., Etarock, M.T., Anangfac, M.T., Nugent, P. and Fondong, V.N., 2008. Role of a geminivirus AV2 protein putative protein kinase C motif on subcellular localization and pathogenicity. Virus Res., 135(1), pp.115-124. https://doi.org/10.1016/j.virusres.2008.02.014

Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J.M., Galzi, S., Filloux, D., Harkins, G.W., Varsani, A. and Martin, D.P., 2018. From spatial metagenomics to molecular characterization of plant viruses: a geminivirus case study. Adv. Virus Res., 101, pp.55-83. https://doi.org/10.1016/bs.aivir.2018.02.003

Clérot, D. and Bernardi, F., 2006. DNA helicase activity is associated with the replication initiator protein rep of tomato yellow leaf curl geminivirus. J. Virol., 80(22), pp.11322-11330. https://doi.org/10.1128/JVI.00924-06

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res., 14(6), pp.1188-1190. https://doi.org/10.1101/gr.849004

Csorba, T., Kontra, L. and Burgyán, J., 2015. Viral silencing suppressors: Tools forged to fine-tune host-pathogen coexistence. Virology, 479, pp.85-103. https://doi.org/10.1016/j.virol.2015.02.028

Cuthbertson, J.M., Doyle, D.A. and Sansom, M.S., 2005. Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng. Des. and Sel., 18(6), pp.295-308. https://doi.org/10.1093/protein/gzi032

Czosnek, H., Ghanim, M. and Ghanim, M., 2002. The circulative pathway of begomoviruses in the whitefly vector Bemisia tabaci—insights from studies with Tomato yellow leaf curl virus. Ann. Appl. Biol., 140(3), pp.215-231. https://doi.org/10.1111/j.1744-7348.2002.tb00175.x

DAFF, 2019. A profile of the South African tomato market value chain. South African Department of Agriculture, Forestry and Fisheries, Pretoria, South Africa, 43 p. https://www.dalrrd.gov.za/doaDev/sideMenu/Marketing/Annual%20Publications/Tomato%20Market%20Value%20Chain%20Profile%202019.pdf

Delatte, H., Martin, D.P., Naze, F., Goldbach, R., Reynaud, B., Peterschmitt, M. and Lett, J.M., 2005. South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. J. Gen. Virol., 86(5), pp.1533-1542. https://doi.org/10.1099/vir.0.80805-0

Deller, M.C., Kong, L. and Rupp, B., 2016. Protein stability: a crystallographer's perspective. Acta Crystallogr. Sect. F Struct. Biol. Commun., 72(2), pp.72-95. https://doi.org/10.1107/S2053230X15024619

Deom, C.M., Brewer, M.T. and Severns, P.M., 2021. Positive selection and intrinsic disorder are associated with multifunctional C4 (AC4) proteins and geminivirus diversification. Sci. Rep., 11, 11150. https://doi.org/10.1038/s41598-021-90557-0

Desbiez, C., David, C., Mettouchi, A., Laufs, J. and Gronenborn, B., 1995. Rep protein of tomato yellow leaf curl geminivirus has an ATPase activity required for viral DNA replication. Proc. Natl. Acad. Sci., 92(12), pp.5640-5644. https://doi.org/10.1073/pnas.92.12.5640

Devendran, R., Namgial, T., Reddy, K.K., Kumar, M., Zarreen, F. and Chakraborty, S., 2022. Insights into the multifunctional roles of geminivirus-encoded proteins in pathogenesis. Arch. Virol., 167, pp.307–326. https://doi.org/10.1007/s00705-021-05338-x

Dosztányi, Z., Sándor, M., Tompa, P. and Simon, I., 2007. Prediction of protein disorder at the domain level. Curr. Protein and Pept. Sci., 8(2), pp.161-171. https://doi.org/10.2174/138920307780363406

Doyle J., 1991. DNA Protocols for Plants. In: Hewitt G.M., Johnston A.W.B., Young J.P.W. (eds) *Molecular Techniques in* Taxonomy. NATO ASI Series (Series H: Cell Biology), vol 57, Springer, Berlin, Heidelberg, pp.283-293. https://doi.org/10.1007/978-3-642-83962-7_18

Du, Z., Chen, M., Wang, Z., Liu, Y., Zhang, S. and He, Z., 2014. Isolation and molecular characterization of a distinct begomovirus and its associated betasatellite infecting Hedyotis uncinella (Hook. et Arn.) in Vietnam. Virus genes, 48(3), pp.557-561. https://doi.org/10.1007/s11262-014-1043-2

Duffaud, G.D., Lehnhardt, S.K., March, P.E. and Inouye, M., 1985. Structure and function of the signal peptide. In Bronner, F. (ed), *Current topics in membranes and transport*, vol. 24, Academic Press, New York, pp.65-104

Dumon-Seignovert, L., Cariot, G. and Vuillard, L., 2004. The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21 (DE3), C41 (DE3), and C43 (DE3). Protein Expr. Purif., 37(1), pp.203-206. https://doi.org/10.1016/j.pep.2004.04.025

Duong-Ly, K.C. and Gabelli, S.B., 2014a. Salting out of proteins using ammonium sulfate precipitation. In: Lorsch, J. (ed.) *Methods in enzymology*, vol 541, Academic Press, New York, pp. 85-94. https://doi.org/10.1016/B978-0-12-420119-4.00007-0

Duong-Ly, K.C. and Gabelli, S.B., 2014b. Gel filtration chromatography (size exclusion chromatography) of proteins. In: Lorsch, J. (ed) *Methods in enzymology*, vol 541, Academic Press, New York, pp. 104-114. https://doi.org/10.1016/B978-0-12-420119-4.00009-4

Eagle, P.A. and Hanley-Bowdoin, L., 1997. cis elements that contribute to geminivirus transcriptional regulation and the efficiency of DNA replication. J. Virol., 71(9), pp.6947-6955. https://doi.org/10.1128/jvi.71.9.6947-6955.1997

Eagle, P.A., Orozco, B.M. and Hanley-Bowdoin, L., 1994. A DNA sequence required for geminivirus replication also mediates transcriptional regulation. Plant Cell, 6(8), pp.1157-1170. https://doi.org/10.1105/tpc.6.8.1157

Egelkrout, E.M., Robertson, D. and Hanley-Bowdoin, L., 2001. Proliferating cell nuclear antigen transcription is repressed through an E2F consensus element and activated by geminivirus infection in mature leaves. Plant Cell, 13(6), pp.1437-1452. https://doi.org/10.1105/TPC.010004

Enemark, E.J. and Joshua-Tor, L., 2006. Mechanism of DNA translocation in a replicative hexameric helicase. Nature, 442(7100), pp.270-275. https://doi.org/10.1038/nature04943

Esmaeili, M., Heydarnejad, J., Massumi, H. and Varsani, A., 2015. Analysis of watermelon chlorotic stunt virus and tomato leaf curl Palampur virus mixed and pseudo-recombination infections. Virus genes, 51(3), pp.408-416. https://doi.org/10.1007/s11262-015-1250-5

Esterhuizen, L.L., 2012. *Ph.D. thesis.* A study of the South African Tomato curly stunt virus pathosystem: Epidemiology, molecular diversity and resistance. University of Johannesburg (South Africa).

Everett, B.A., Litzau, L.A., Tompkins, K., Shi, K., Nelson, A., Aihara, H., Evans Iii, R.L. and Gordon, W.R., 2019. Crystal structure of the Wheat dwarf virus Rep domain. Acta Crystallogr. Sect. F Struct. Biol. Commun., 75(12), pp.744-749. https://doi.org/10.1107/S2053230X19015796

FAOSTAT, 2022. Food and Agriculture Organisation Statistics Database. Food and Agriculture Organisation of the United Nations, Rome, Italy. https://www.fao.org/faostat/en/

Farooq, T., Liu, D., Zhou, X. and Yang, Q., 2019. Tomato Yellow Leaf Curl China Virus Impairs Photosynthesis in the Infected *Nicotiana benthamiana* with βC1 as an Aggravating Factor. Plant Pathol. J., 35(5), pp.521-529. https://dx.doi.org/10.5423%2FPPJ.OA.04.2019.0120

Fauquet, C.M., Briddon, R.W., Brown, J.K., Moriones, E., Stanley, J., Zerbini, M. and Zhou, X., 2008. Geminivirus strain demarcation and nomenclature. Arch. Virol., 153(4), pp.783-821. https://doi.org/10.1007/s00705-008-0037-6

Ferro, C.G., Zerbini, F.M., Navas-Castillo, J. and Fiallo-Olivé, E., 2021. Revealing the Complexity of Sweepovirus-Deltasatellite–Plant Host Interactions: Expanded Natural and Experimental Helper Virus Range and Effect Dependence on Virus-Host Combination. Microorganisms, 9(5), p.1018. https://doi.org/10.3390/microorganisms9051018

Ferron, F., Longhi, S., Canard, B. and Karlin, D., 2006. A practical overview of protein disorder prediction methods. Proteins, 65(1), pp.1-14. https://doi.org/10.1002/prot.21075

Fiallo-Olivé, E., Lett, J.M., Martin, D.P., Roumagnac, P., Varsani, A., Zerbini, F.M. and Navas-Castillo, J., 2021. ICTV virus taxonomy profile: geminiviridae 2021. J. Gen. Virol., 102(12), p.001696. https://doi.org/10.1099/jgv.0.001696

Firth, A.E. and Brierley, I., 2012. Non-canonical translation in RNA viruses. J. Gen. Virol., 93(Pt 7), pp.1385-1409. https://dx.doi.org/10.1099%2Fvir.0.042499-0

Fomenkov, A., Sun, Z., Dila, D.K., Anton, B.P., Roberts, R.J. and Raleigh, E.A., 2017. EcoBLMcrX, a classical modification-dependent restriction enzyme in Escherichia coli B: characterization in vivo and in vitro with a new approach to cleavage site determination. PloS one, 12(6), p.e0179853. https://doi.org/10.1371/journal.pone.0179853

Fondong, V.N., 2013. Geminivirus protein structure and function. Mol. Plant Pathol., 14(6), pp.635-649. https://doi.org/10.1111/mpp.12032

Fondong, V.N., 2019. The ever-expanding role of C4/AC4 in geminivirus infection: punching above its weight?. Mol. Plant, 12(2), pp.145-147. https://doi.org/10.1016/j.molp.2018.12.006

Fontenelle, M.R., Luz, D.F., Gomes, A.P.S., Florentino, L.H., Zerbini, F.M. and Fontes, E.P., 2007. Functional analysis of the naturally recombinant DNA-A of the bipartite begomovirus Tomato chlorotic mottle virus. Virus Res., 126(1-2), pp.262-267. https://doi.org/10.1016/j.virusres.2007.02.009

Fontes, E.P., Eagle, P.A., Sipe, P.S., Luckow, V.A. and Hanley-Bowdoin, L., 1994. Interaction between a geminivirus replication protein and origin DNA is essential for viral replication. J. Biol. Chem., 269(11), pp.8459-8465. https://doi.org/10.1016/S0021-9258(17)37216-2

Francis, D.M. and Page, R., 2010. Strategies to optimize protein expression in E. coli. Curr. Protoc. Protein Sci., 61(1), pp.5-24. https://doi.org/10.1002/0471140864.ps0524s61

Fukunaga, R. and Doudna, J.A., 2009. dsRNA with 5′ overhangs contributes to endogenous and antiviral RNA silencing pathways in plants. EMBO J., 28(5), pp.545-555. https://doi.org/10.1038/emboj.2009.2

Gafni, Y. and Epel, B.L., 2002. The role of host and viral proteins in intra-and inter-cellular trafficking of geminiviruses. Physiol. Mol. Plant Pathol., 60(5), pp.231-241. https://doi.org/10.1006/pmpp.2002.0402

Gai, D., Zhao, R., Li, D., Finkielstein, C.V. and Chen, X.S., 2004. Mechanisms of conformational change for a replicative hexameric helicase of SV40 large tumor antigen. Cell, 119(1), pp.47-60. https://doi.org/10.1016/j.cell.2004.09.017

Gamage, D.G., Gunaratne, A., Periyannan, G.R. and Russell, T.G., 2019. Applicability of instability index for in vitro protein stability prediction. Protein Pept. Lett., 26(5), pp.339-347. https://doi.org/10.2174/0929866526666190228144219

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D. and Bairoch, A., 2005. Protein identification and analysis tools on the ExPASy server. In: J.M. Walker, (ed), *The Proteomics Protocols Handbook*. Humana Press, Totowa, pp.571-607. https://doi.org/10.1385/1-59259-890-0:571

George, B., Ruhel, R., Mazumder, M., Sharma, V.K., Jain, S.K., Gourinath, S. and Chakraborty, S., 2014. Mutational analysis of the helicase domain of a replication initiator protein reveals critical roles of Lys 272 of the B′ motif and Lys 289 of the β-hairpin loop in geminivirus replication. J. Gen. Virol., 95(7), pp.1591-1602. https://doi.org/10.1099/vir.0.064923-0

Glick, E., Zrachya, A., Levy, Y., Mett, A., Gidoni, D., Belausov, E., Citovsky, V. and Gafni, Y., 2008. Interaction with host SGS3 is required for suppression of RNA

silencing by tomato yellow leaf curl virus V2 protein. Proc. Acad. Natl. Sci., 105(1), pp.157-161. https://dx.doi.org/10.1073%2Fpnas.0709036105

Gnanasekaran, P., KishoreKumar, R., Bhattacharyya, D., Vinoth Kumar, R. and Chakraborty, S., 2019. Multifaceted role of geminivirus associated betasatellite in pathogenesis. Mol. Plant Pathol., 20(7), pp.1019-1033. https://doi.org/10.1111/mpp.12800

Gong, P., Tan, H., Zhao, S., Li, H., Liu, H., Ma, Y., Zhang, X., Rong, J., Fu, X., Lozano-Duran, R. and Li, F., 2021. Geminiviral genomes encode additional proteins with specific subcellular localizations and virulence function. Nat. Commun., 12, 4278. https://doi.org/10.1038/s41467-021-24617-4

Goodin, M.M., Zaitlin, D., Naidu, R.A. and Lommel, S.A., 2008. Nicotiana benthamiana: its history and future as a model for plant–pathogen interactions. Mol. Plant Microbe Interact., 21(8), pp.1015-1026. https://doi.org/10.1094/MPMI-21-8-1015

Gover, O., Peretz, Y., Mozes-Koch, R., Maori, E., Rabinowitch, H.D. and Sela, I., 2014. Only minimal regions of tomato yellow leaf curl virus (TYLCV) are required for replication, expression and movement. Arch. Virol., 159(9), pp.2263-2274. https://doi.org/10.1007/s00705-014-2066-7

Green, M.R. and Sambrook, J., 2020. The Inoue method for preparation and transformation of competent Escherichia coli:“Ultracompetent” cells. Cold Spring Harb. Protoc., 2020(6), pp.pdb-prot101196. https://doi.org/10.1101/2020.06.26.173526v2

Guo, X. and Huang, L., 2010. A superfamily 3 DNA helicase encoded by plasmid pSSVi from the hyperthermophilic archaeon Sulfolobus solfataricus unwinds DNA as a higher-order oligomer and interacts with host primase. J. Bacteriol., 192(7), pp.1853-1864. https://doi.org/10.1128/JB.01300-09

Guruprasad, K., Reddy, B.B. and Pandit, M.W., 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. Des. Sel., 4(2), pp.155-161. https://doi.org/10.1093/protein/4.2.155

Gutierrez, C., 2000. DNA replication and cell cycle in plants: learning from geminiviruses. The EMBO J., 19(5), pp.792-799. https://doi.org/10.1093/emboj/19.5.792

Gutierrez, C., Ramirez-Parra, E., Castellano, M.M., Sanz-Burgos, A.P., Luque, A. and Missich, R., 2004. Geminivirus DNA replication and cell cycle interactions. Vet. Microbiol., 98(2), pp.111-119. https://doi.org/10.1016/j.vetmic.2003.10.012

Hak, H., Levy, Y., Chandran, S.A., Belausov, E., Loyter, A., Lapidot, M. and Gafni, Y., 2015. TYLCV-Is movement in planta does not require V2 protein. Virology, 477, pp.56-60. https://doi.org/10.1016/J.VIROL.2015.01.007

Haley, A., Zhan, X., Richardson, K., Head, K. and Morris, B., 1992. Regulation of the activities of African cassava mosaic virus promoters by the AC1, AC2, and AC3 gene products. Virology, 188(2), pp.905-909. https://doi.org/10.1016/0042-6822(92)90551-Y

Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D. and Mansoor, S., 2013. Geminiviruses: masters at redirecting and reprogramming plant processes. Nat. Rev. Microbiol., 11(11), pp.777-788. https://doi.org/10.1038/nrmicro3117

Hanley-Bowdoin, L., Settlage, S.B., Orozco, B.M., Nagar, S. and Robertson, D., 2000. Geminiviruses: models for plant DNA replication, transcription, and cell cycle regulation. Crit. Rev. Biochem. Mol. Biol., 35(2), pp.105-140.

Hartitz, M.D., Sunter, G. and Bisaro, D.M., 1999. The tomato golden mosaic virus transactivator (TrAP) is a single-stranded DNA and zinc-binding phosphoprotein with an acidic activation domain. Virology, 263(1), pp.1-14. https://doi.org/10.1006/viro.1999.9925

Hehnle, S., Wege, C. and Jeske, H., 2004. Interaction of DNA with the movement proteins of geminiviruses revisited. J. Virol., 78(14), pp.7698-7706. https://doi.org/10.1128/JVI.78.14.7698-7706.2004

Hesketh, E.L., Saunders, K., Fisher, C., Potze, J., Stanley, J., Lomonossoff, G.P. and Ranson, N.A., 2018. The 3.3 Å structure of a plant geminivirus using cryo-EM. Nat. Commun., 9, 2369.

Heuvelink, E. and Dorais, M., 2005. Crop growth and yield. In: Heuvelink, E. (ed), *Tomatoes*, CAB International, London, pp.104-120.

Hickman, A.B. and Dyda, F., 2005. Binding and unwinding: SF3 viral helicases. Curr. Opin. Struct. Biol., 15(1), pp.77-85. https://doi.org/10.1016/j.sbi.2004.12.001

Hidayat, S.H., Chatchawankanpanich, O. and Aidawati, N., 2008. Molecular identification and sequence analysis of tobacco leaf curl begomovirus from Jember, East Java, Indonesia. Hayati, 15(1), pp.13-17. https://doi.org/10.4308/hjb.15.1.13

Hipp, K., Grimm, C., Jeske, H. and Böttcher, B., 2017. Near-atomic resolution structure of a plant geminivirus determined by electron cryomicroscopy. Structure, 25(8), pp.1303-1309. https://doi.org/10.1016/j.str.2017.06.013

Hoffer, P., Ivashuta, S., Pontes, O., Vitins, A., Pikaard, C., Mroczka, A., Wagner, N. and Voelker, T., 2011. Posttranscriptional gene silencing in nuclei. Proc. Natl. Acad. Sci., 108(1), pp.409-414. https://doi.org/10.1073/pnas.1009805108

Hoffman, B.J., Broadwater, J.A., Johnson, P., Harper, J., Fox, B.G. and Kenealy, W.R., 1995. Lactose fed-batch overexpression of recombinant metalloproteins in Escherichia coli BL21 (DE3): process control yielding high levels of metal-incorporated, soluble protein. Protein Expr. Purif., 6(5), pp.646-654. https://doi.org/10.1006/prep.1995.1085

Höfgen, R. and Willmitzer, L., 1988. Storage of competent cells for Agrobacterium transformation. Nucleic Acids Res., 16(20), p.9877. https://dx.doi.org/10.1093%2Fnar%2F16.20.9877

Hong, Y.G., Robinson, D.J. and Harrison, B.D., 1993. Nucleotide sequence evidence for the occurrence of three distinct whitefly-transmitted geminiviruses in cassava. J. Gen. Virol., 74(11), pp.2437-2443. https://doi.org/10.1099/0022-1317-74-11-2437

Hong, Y. and Stanley, J., 1995. Regulation of African cassava mosaic virus complementary-sense gene expression by N-terminal sequences of the replication-associated protein AC1. J. Gen. Virol., 76(10), pp.2415-2422. https://doi.org/10.1099/0022-1317-76-10-2415

Hu, T., Song, Y., Wang, Y. and Zhou, X., 2020. Functional analysis of a novel βV1 gene identified in a geminivirus betasatellite. Sci. China Life Sci., 63(5), pp.688-696. https://doi.org/10.1007/s11427-020-1654-x

Hu, S., Wu, Z., Yang, L. and Fung, M.C., 2009. Molecular cloning and expression of a functional anti-inflammatory protein, Sj16, of Schistosoma japonicum. Int. J. parasitol., 39(2), pp.191-200. https://doi.org/10.1016/j.ijpara.2008.06.017

Hunt, I., 2005. From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. Protein Expr. Purif., 40(1), pp.1-22. https://doi.org/10.1016/j.pep.2004.10.018

Hur, J., Choi, E., Buckley, K.J., Lee, S. and Davis, K.R., 2008. Identification of a Promoter Motif Involved in Curtovirus Sense-Gene Expression in Transgenic Arabidopsis. Molec. Cells, 26(2), pp.131-139.

Ilyina, T.V. and Koonin, E.V., 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res., 20(13), pp.3279-3285. https://doi.org/10.1093/nar/20.13.3279

Jacobson, A.L., Duffy, S. and Sseruwagi, P., 2018. Whitefly-transmitted viruses threatening cassava production in Africa. Curr. Opin. Virol., 33, pp.167-176. https://doi.org/10.1016/j.coviro.2018.08.016

Jahandideh, S., Jaroszewski, L. and Godzik, A., 2014. Improving the chances of successful protein structure determination with a random forest classifier. Acta Crystallogr. Sect. D Struct. Biol., 70(3), pp.627-635. https://doi.org/10.1107/S1399004713032070

Jain, A. and Chugh, A., 2016. Mitochondrial transit peptide exhibits cell penetration ability and efficiently delivers macromolecules to mitochondria. FEBS Lett., 590(17), pp.2896-2905. https://doi.org/10.1002/1873-3468.12329

James, J.A., Escalante, C.R., Yoon-Robarts, M., Edwards, T.A., Linden, R.M. and Aggarwal, A.K., 2003. Crystal structure of the SF3 helicase from adeno-associated virus type 2. Structure, 11(8), pp.1025-1035. https://doi.org/10.1016/S0969-2126(03)00152-7

James, J.A., Aggarwal, A.K., Linden, R.M. and Escalante, C.R., 2004. Structure of adeno-associated virus type 2 Rep40–ADP complex: insight into nucleotide recognition and catalysis by superfamily 3 helicases. Proc. Natl. Acad. Sci., 101(34), pp.12455-12460. https://doi.org/10.1073/pnas.0403454101

Javed, A., Major, B., Stead, J.A., Sanders, C.M. and Orlova, E.V., 2021. Unwinding of a DNA replication fork by a hexameric viral helicase. Nat. Commun., 12, 5535. https://doi.org/10.1038/s41467-021-25843-6

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292(2), pp.195-202. https://doi.org/10.1006/jmbi.1999.3091

Jones, J. B., 2007. Field production in soil. In: *Tomato plant culture: In the field, greenhouse and home garden*, 2nd ed, CRC Press, Boca Raton, pp.179-180.

Jones, D.T. and Cozzetto, D., 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics, 31(6), pp.857-863. https://doi.org/10.1093/bioinformatics/btu744

Jupin, I., Hericourt, F., Benz, B. and Gronenborn, B., 1995. DNA replication specificity of TYLCV geminivirus is mediated by the amino-terminal 116 amino acids of the Rep protein. FEBS Lett., 362(2), pp.116-120. https://doi.org/10.1016/0014-5793(95)00221-T

Kaliappan, K., Choudhury, N.R., Suyal, G. and Mukherjee, S.K., 2012. A novel role for RAD54: this host protein modulates geminiviral DNA replication. The FASEB Journal, 26(3), pp.1142-1160. https://doi.org/10.1096/fj.11-188508

Käll, L., Krogh, A. and Sonnhammer, E.L., 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res., 35(suppl_2), pp.W429-W432. https://doi.org/10.1093/nar/gkm256

Kamarudin, N.H.A., Rahman, R.N.Z.R.A., Ali, M.S.M., Leow, T.C., Basri, M. and Salleh, A.B., 2014. Unscrambling the effect of C-terminal tail deletion on the stability of a cold-adapted, organic solvent stable lipase from Staphylococcus epidermidis AT2. Mol. Biotechnol., 56(8), pp.747-757. https://doi.org/10.1007/s12033-014-9753-1

Kanakala, S., Jyothsna, P., Shukla, R., Tiwari, N., Veer, B.S., Swarnalatha, P., Krishnareddy, M. and Malathi, V.G., 2013. Asymmetric synergism and heteroencapsidation between two bipartite begomoviruses, tomato leaf curl New Delhi virus and tomato leaf curl Palampur virus. Virus Res., 174(1-2), pp.126-136. https://doi.org/10.1016/j.virusres.2013.03.011

Kazlauskas, D., Varsani, A., Koonin, E.V. and Krupovic, M., 2019. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat. Commun., 10, 2435. https://doi.org/10.1038/s41467-019-11433-0

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc., 10(6), pp.845-858. https://doi.org/10.1038/nprot.2015.053

Khan, Z.A., Abdin, M.Z. and Khan, J.A., 2015. Functional characterization of a strong bi-directional constitutive plant promoter isolated from cotton leaf curl Burewala virus. PLoS One, 10(3), p.e0121656. https://doi.org/10.1371/journal.pone.0121656

Kheyr-Pour, A., Bananej, K., Dafalla, G.A., Caciagli, P., Noris, E., Ahoonmanesh, A., Lecoq, H. and Gronenborn, B., 2000. Watermelon chlorotic stunt virus from the Sudan and Iran: sequence comparisons and identification of a whitefly-transmission determinant. Phytopathology, 90(6), pp.629-635. https://doi.org/10.1094/PHYTO.2000.90.6.629

Kielkopf, C.L., Bauer, W. and Urbatsch, I.L., 2020. Purification of polyhistidine-tagged proteins by immobilized metal affinity chromatography. Cold Spring Harb. Protoc., 2020(6), pp.pdb-prot102194. https://doi.org/10.1101/pdb.prot102194

Kimple, M.E., Brill, A.L. and Pasker, R.L., 2013. Overview of affinity tags for protein purification. Curr. Protoc. Protein Sci., 73(1), pp.9-9. https://doi.org/10.1002/0471140864.ps0909s73

Kittelmann, K., Rau, P., Gronenborn, B. and Jeske, H., 2009. Plant geminivirus Rep protein induces rereplication in fission yeast. J. Virol., 83(13), pp.6769-6778. https://doi.org/10.1128/JVI.02491-08

Kon, T., Sharma, P. and Ikegami, M., 2007. Suppressor of RNA silencing encoded by the monopartite tomato leaf curl Java begomovirus. Arch. Virol., 152(7), pp.1273-1282. https://doi.org/10.1007/s00705-007-0957-6

Kong, L.J. and Hanley-Bowdoin, L., 2002. A geminivirus replication protein interacts with a protein kinase and a motor protein that display different expression patterns during plant development and infection. Plant Cell, 14(8), pp.1817-1832. https://doi.org/10.1105/tpc.003681

Kong, L.J., Orozco, B.M., Roe, J.L., Nagar, S., Ou, S., Feiler, H.S., Durfee, T., Miller, A.B., Gruissem, W., Robertson, D. and Hanley-Bowdoin, L., 2000. A geminivirus replication protein interacts with the retinoblastoma protein through a novel domain to determine symptoms and tissue specificity of infection in plants. EMBO J., 19(13), pp.3485-3495. https://doi.org/10.1093/emboj/19.13.3485

Koonin, E.V., 1993. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. Nucleic Acids Res., 21(11), pp.2541-2547. https://doi.org/10.1093/nar/21.11.2541

Koonin, E.V. and Ilyina, T.V., 1992. Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J. Gen. Virol., 73(10), pp.2763-2766. https://doi.org/10.1099/0022-1317-73-10-2763

Kõressaar, T., Lepamets, M., Kaplinski, L., Raime, K., Andreson, R. and Remm, M., 2018. Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics, 34(11), pp.1937-1938. https://doi.org/10.1093/bioinformatics/bty036

Kosugi, S., Hasebe, M., Tomita, M. and Yanagawa, H., 2009. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. Proc. Natl. Acad. of Sci., 106(25), pp.10171-10176. https://doi.org/10.1073/pnas.0900604106

Kramer, R.M., Shende, V.R., Motl, N., Pace, C.N. and Scholtz, J.M., 2012. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. Biophys. J., 102(8), pp.1907-1915. https://dx.doi.org/10.1016%2Fj.bpj.2012.01.060

Krenz, B., Neugart, F., Kleinow, T. and Jeske, H., 2011. Self-interaction of Abutilon mosaic virus replication initiator protein (Rep) in plant cell nuclei. Virus Res., 161(2), pp.194-197. https://doi.org/10.1016/j.virusres.2011.07.020

Krogh, A., Larsson, B., Von Heijne, G. and Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol., 305(3), pp.567-580. https://doi.org/10.1006/jmbi.2000.4315

Kühlbrandt, W., 2014. Microscopy: cryo-EM enters a new era. elife, 3, p.e03678. https://doi.org/10.7554/elife.03678

Kumar, R.V., 2019. Plant antiviral immunity against geminiviruses and viral counter-defense for survival. Front. Microbiol., 10, 1460. https://doi.org/10.3389/fmicb.2019.01460

Kumar, R. V. and Shivaprasad, P.V., 2020. Plant-virus-insect tritrophic interactions: insights into the functions of geminivirus virion-sense strand genes. Proc. Royal Soc. B, 287(1936), p.20201846. https://doi.org/10.1098/rspb.2020.1846

Kushwaha, N.K., Bhardwaj, M. and Chakraborty, S., 2017. The replication initiator protein of a geminivirus interacts with host monoubiquitination machinery and stimulates transcription of the viral genome. PLoS Pathog., 13(8), p.e1006587. https://doi.org/10.1371/journal.ppat.1006587

Laible, M. and Boonrod, K., 2009. Homemade site directed mutagenesis of whole plasmids. Journal of visualized experiments: JoVE, (27). https://dx.doi.org/10.3791%2F1135

Lapidot, M., Ben-Joseph, R., Cohen, L., Machbash, Z. and Levy, D., 2006. Development of a scale for evaluation of Tomato yellow leaf curl virus resistance level in tomato plants. Phytopathology, 96(12), pp.1404-1408. https://doi.org/10.1094/PHYTO-96-1404

Lapidot, M., Friedmann, M., Pilowsky, M., Ben-Joseph, R. and Cohen, S., 2001. Effect of host plant resistance to Tomato yellow leaf curl virus (TYLCV) on virus acquisition and transmission by its whitefly vector. Phytopathology, 91(12), pp.1209-1213. https://doi.org/10.1094/PHYTO.2001.91.12.1209

Laufs, J., Jupin, I., David, C., Schumacher, S., Heyraud-Nitschke, F. and Gronenborn, B., 1995a. Geminivirus replication: genetic and biochemical characterization of Rep protein function, a review. Biochimie, 77(10), pp.765-773. https://doi.org/10.1016/0300-9084(96)88194-6

Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S.G., Schell, J. and Gronenborn, B., 1995b. In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. Proc. Natl. Acad. Sci., 92(9), pp.3879-3883. https://doi.org/10.1073/pnas.92.9.3879

Laufs, J., Schumacher, S., Geisler, N., Jupin, I. and Gronenborn, B., 1995c. Identification of the nicking tyrosine of geminivirus Rep protein. FEBS Lett., 377(2), pp.258-262. https://doi.org/10.1016/0014-5793(95)01355-5

Lawry, R., Martin, D.P., Shepherd, D.N., van Antwerpen, T. and Varsani, A., 2009. A novel sugarcane-infecting mastrevirus from South Africa. Arch. Virol., 154(10), pp.1699-1703. https://doi.org/10.1007/s00705-009-0490-x

Lazarowitz, S.G., 1988. Infectivity and complete nucleotide sequence of the genome of a South African isolate of maize streak virus. Nucleic Acids Res., 16(1), pp.229-249. https://doi.org/10.1093/nar/16.1.229

Lazarowitz, S.G., Wu, L.C., Rogers, S.G. and Elmer, J.S., 1992. Sequence-specific interaction with the viral AL1 protein identifies a geminivirus DNA replication origin. Plant Cell, 4(7), pp.799-809. https://doi.org/10.1105/tpc.4.7.799

Lefeuvre, P., Martin, D.P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. and Lett, J.M., 2007a. Begomovirus 'melting pot'in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. J. Gen. Virol., 88(12), pp.3458-3468. https://doi.org/10.1099/vir.0.83252-0

Lefeuvre, P., Delatte, H., Naze, F., Dogley, W., Reynaud, B. and Lett, J.M., 2007b. A new tomato leaf curl virus from the Seychelles archipelago. Plant Pathol., 56(2), pp.342-342. https://doi.org/10.1111/j.1365-3059.2006.01471.x

Legg, J., Somado, E.A., Barker, I., Beach, L., Ceballos, H., Cuellar, W., Elkhoury, W., Gerling, D., Helsen, J., Hershey, C. and Jarvis, A., 2014. A global alliance declaring

war on cassava viruses in Africa. Food Secur., 6(2), pp.231-248. https://doi.org/10.1007/s12571-014-0340-x

Leke, W.N., Mignouna, D.B., Brown, J.K. and Kvarnheden, A., 2015. Begomovirus disease complex: emerging threat to vegetable production systems of West and Central Africa. Agric. Food Secur., 4(1), pp.1-14. https://doi.org/10.1186/s40066-014-0020-2

Li, F., Xu, X., Huang, C., Gu, Z., Cao, L., Hu, T., Ding, M., Li, Z. and Zhou, X., 2015. The AC 5 protein encoded by Mungbean yellow mosaic India virus is a pathogenicity determinant that suppresses RNA silencing-based antiviral defenses. New Phytol., 208(2), pp.555-569. https://doi.org/10.1111/nph.13473

Li, M., Li, C., Jiang, K., Li, K., Zhang, J., Sun, M., Wu, G. and Qing, L., 2021a. Characterization of Pathogenicity-Associated V2 Protein of Tobacco Curly Shoot Virus. International J. Mol. Sci., 22(2), p.923. https://doi.org/10.3390/ijms22020923

Li, P., Su, F., Meng, Q., Yu, H., Wu, G., Li, M. and Qing, L., 2021b. The C5 protein encoded by Ageratum leaf curl Sichuan virus is a virulence factor and contributes to the virus infection. Mol. Plant Pathol., 22(9), pp.1149-1158. https://doi.org/10.1111/mpp.13103

Lichty, J.J., Malecki, J.L., Agnew, H.D., Michelson-Horowitz, D.J. and Tan, S., 2005. Comparison of affinity tags for protein purification. Protein Expr. Purif., 41(1), pp.98-105. https://doi.org/10.1016/j.pep.2005.01.019

Liu, H. and Naismith, J.H., 2008. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. BMC Biotechnol., 8(1), pp.1-10. https://doi.org/10.1186/1472-6750-8-91

Liu, S.R., Zhou, J.J., Hu, C.G., Wei, C.L. and Zhang, J.Z., 2017. MicroRNA-mediated gene silencing in plant defense and viral counter-defense. Front. Microbiol., 8, p.1801. https://doi.org/10.3389/fmicb.2017.01801

Liu, D., Shi, L., Han, C., Yu, J., Li, D. and Zhang, Y., 2012. Validation of Reference Genes for Gene Expression Studies in Virus-Infected Nicotiana benthamiana Using Quantitative Real-Time PCR. PLoS ONE, 7(9), pp.e46451-e46451. https://doi.org/10.1371/journal.pone.0046451

Liu, L., van Tonder, T., Pietersen, G., Davies, J.W. and Stanley, J., 1997. Molecular characterization of a subgroup I geminivirus from a legume in South Africa. J. Gen. Virol., 78(8), pp.2113-2117. https://doi.org/10.1099/0022-1317-78-8-2113

Liu, S.R., Zhou, J.J., Hu, C., Wei, C. and Zhang, J.Z., 2017. MicroRNA-mediated gene silencing in plant defense and viral counter-defense. Front. Microbiol., 8, 1801. https://doi.org/10.3389/fmicb.2017.01801

Londoño, A., Riego-Ruiz, L. and Argüello-Astorga, G.R., 2010. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch. Virol., 155(7), pp.1033-1046. https://doi.org/10.1007/s00705-010-0674-4

Lozano, G., Trenado, H.P., Fiallo-Olivé, E., Chirinos, D., Geraud-Pouey, F., Briddon, R.W. and Navas-Castillo, J., 2016. Characterization of non-coding DNA satellites associated with sweepoviruses (genus Begomovirus, Geminiviridae)–definition of a distinct class of begomovirus-associated satellites. Frontiers in Microbiology, 7, p.162. https://doi.org/10.3389/fmicb.2016.00162

Luna, A.P., Rodríguez-Negrete, E.A., Morilla, G., Wang, L., Lozano-Durán, R., Castillo, A.G. and Bejarano, E.R., 2017. V2 from a curtovirus is a suppressor of post-transcriptional gene silencing. J. Gen. Virol., 98(10), pp.2607-2614. https://doi.org/10.1099/jgv.0.000933

Luna, A.P., Romero-Rodríguez, B., Rosas-Díaz, T., Cerero, L., Rodríguez-Negrete, E.A., Castillo, A.G. and Bejarano, E.R., 2020. Characterization of Curtovirus V2 protein, a functional homolog of begomovirus V2. Front. Plant Sci., 11, 835. https://doi.org/10.3389/fpls.2020.00835

Luque, A., Sanz-Burgos, A.P., Ramirez-Parra, E., Castellano, M.M. and Gutierrez, C., 2002. Interaction of geminivirus Rep protein with replication factor C and its potential role during geminivirus DNA replication. Virology, 302(1), pp.83-94. https://doi.org/10.1006/viro.2002.1599

Mabvakure, B., Martin, D.P., Kraberger, S., Cloete, L., van Brunschot, S., Geering, A.D., Thomas, J.E., Bananej, K., Lett, J.M., Lefeuvre, P. and Varsani, A., 2016. Ongoing geographical spread of Tomato yellow leaf curl virus. Virology, 498, pp.257-264. https://doi.org/10.1016/j.virol.2016.08.033

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D. and Lopez, R., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res., 47(W1), pp.W636-W641. https://doi.org/10.1093/nar/gkz268

Mansoor, S., Khan, S.H., Bashir, A., Saeed, M., Zafar, Y., Malik, K.A., Briddon, R., Stanley, J. and Markham, P.G., 1999. Identification of a novel circular single-stranded DNA associated with cotton leaf curl disease in Pakistan. Virology, 259(1), pp.190-199. https://doi.org/10.1006/viro.1999.9766

McPherson, A. and Gavira, J.A., 2014. Introduction to protein crystallization. Acta Crystallogr. Sect. F Struct. Biol. Commun., 70(1), pp.2-20. https://dx.doi.org/10.1107%2FS2053230X13033141

Medina-Puche, L., Orílio, A.F., Zerbini, F.M. and Lozano-Durán, R., 2021. Small but mighty: Functional landscape of the versatile geminivirus-encoded C4 protein. PLoS Pathog., 17(10), p.e1009915. https://doi.org/10.1371/journal.ppat.1009915

Melgarejo, T.A., Kon, T., Rojas, M.R., Paz-Carrasco, L., Zerbini, F.M. and Gilbertson, R.L., 2013. Characterization of a new world monopartite begomovirus causing leaf curl disease of tomato in Ecuador and Peru reveals a new direction in geminivirus evolution. J. Virol., 87(10), pp.5397-5413. https://doi.org/10.1128/JVI.00234-13

Miozzi, L., Pantaleo, V., Burgyán, J., Accotto, G.P. and Noris, E., 2013. Analysis of small RNAs derived from tomato yellow leaf curl Sardinia virus reveals a cross reaction between the major viral hotspot and the plant host genome. Virus Res., 178(2), pp.287-296. https://doi.org/10.1016/j.virusres.2013.09.029

Miras, M., Miller, W.A., Truniger, V. and Aranda, M.A., 2017. Non-canonical translation in plant RNA viruses. Front. Plant Sci., 8, p.494. https://doi.org/10.3389/fpls.2017.00494

Miroux, B. and Walker, J.E., 1996. Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J. Mol. Biol., 260(3), pp.289-298. https://doi.org/10.1006/jmbi.1996.0399

Missich, R., Ramirez-Parra, E. and Gutierrez, C., 2000. Relationship of oligomerization to DNA binding of wheat dwarf virus RepA and Rep proteins. Virology, 273(1), pp.178-188. https://doi.org/10.1006/viro.2000.0412

Mivedor, A.S., Pita, J.S., Adjata, K.D., Duclercq, J. and Gumedzoe, Y.M.D., 2017. Genetic Diversity of Tomato (Solanum lycopersicum L.) Begomovirus in Togo. Agric. Sci., 8(12), pp.1402-1414. https://doi.org/10.4236/as.2017.812100

Mohammed, H.S., El Siddig, M.A., El Hussein, A.A., Navas-Castillo, J. and Fiallo-Olivé, E., 2018. Complete genome sequence of datura leaf curl virus, a novel begomovirus infecting Datura innoxia in Sudan, related to begomoviruses causing tomato yellow leaf curl disease. Arch. Virol., 163(1), pp.273-275. https://doi.org/10.1007/s00705-017-3574-z

Moodley, V., Gubba, A. and Mafongoya, P.L., 2019. A survey of whitefly-transmitted viruses on tomato crops in South Africa. Crop Prot., 123, pp.21-29. https://doi.org/10.1016/j.cropro.2019.05.018

Moreau, M.J., McGeoch, A.T., Lowe, A.R., Itzhaki, L.S. and Bell, S.D., 2007. ATPase site architecture and helicase mechanism of an archaeal MCM. Mol. Cell, 28(2), pp.304-314. https://doi.org/10.1016/j.molcel.2007.08.013

Morilla, G., Krenz, B., Jeske, H., Bejarano, E.R. and Wege, C., 2004. Tete a tete of Tomato yellow leaf curl virus and Tomato yellow leaf curl Sardinia virus in single nuclei. J. Virol., 78(19), pp.10715-10723. https://doi.org/10.1128/JVI.78.19.10715-10723.2004

Moriones, E., García-Andrés, S. and Navas-Castillo, J., 2007. Recombination in the TYLCV complex: a mechanism to increase genetic diversity. Implications for plant resistance development. In Czosnek, H. (ed.) Tomato Yellow Leaf Curl Virus Disease, Springer, The Netherlands, pp.119-138.

Mubin, M., Amin, I., Amrao, L., Briddon, R.W. and Mansoor, S., 2010. The hypersensitive response induced by the V2 protein of a monopartite begomovirus is countered by the C2 protein. Mol. Plant Pathol., 11(2), pp.245-254. https://doi.org/10.1111/j.1364-3703.2009.00601.x

Mubin, M., Briddon, R.W. and Mansoor, S., 2019. The V2 protein encoded by a monopartite begomovirus is a suppressor of both post-transcriptional and transcriptional gene silencing activity. Gene, 686, pp.43-48. https://doi.org/10.1016/j.gene.2018.11.002

Mullineaux, P.M., Rigden, J.E., Dry, I.B., Krake, L.R. and Rezaian, M.A., 1993. Mapping of the polycistronic RNAs of tomato leaf curl geminivirus. Virology, 193(1), pp.414-423. https://doi.org/10.1006/viro.1993.1138

Naika, S., van Lidt de Jeude, J., de Goffau, M., Hilmi, M. and van Dam, B., 2005. A brief description of tomato. In van Dam, B. (ed), *Agrodok 17: Cultivation of tomato: production, processing and marketing*, 4th ed, Agromisa Foundation and CTA, Wageningen, pp.6-9.

Nash, T.E., Dallas, M.B., Reyes, M.I., Buhrman, G.K., Ascencio-Ibanez, J.T. and Hanley-Bowdoin, L., 2011. Functional analysis of a novel motif conserved across geminivirus Rep proteins. J. Virol., 85(3), pp.1182-1192. https://doi.org/10.1128/JVI.02143-10

Nawaz-ul-Rehman, M.S. and Fauquet, C.M., 2009. Evolution of geminiviruses and their satellites. FEBS Lett., 583(12), pp.1825-1832. https://doi.org/10.1016/j.febslet.2009.05.045

Nawaz-ul-Rehman, M.S., Nahid, N., Hassan, M. and Mubin, M., 2021. Betasatellites and Deltasatelliles (Tolecusatellitidae). In: Bamford D., Zuckerman, M. (eds) *Encyclopedia of Virology*, vol 3, 4th ed, pp. 239-246. Academic Press, New York. https://doi.org/ 10.1016/B978-0-12-809633-8.21562-8

Nonnecke, I. L., 1989. Tomato. In: *Vegetable Production*, Van Nostrand Reinhold, New York, pp.200-203.

Noris, E., Jupin, I., Accotto, G.P. and Gronenborn, B., 1996. DNA-binding activity of the C2 protein of tomato yellow leaf curl geminivirus. Virology, 217(2), pp.607-612. https://doi.org/10.1006/viro.1996.0157

Odame, H., Musyoka, P. and Kere, J., 2009. Kenya: Maize, tomato and dairy. In Larsen, K., Kim, R. and Theus, F. (eds.) *Agribusiness and Innovation Systems in Africa*, The World Bank, Washington, pp.89-134.

Ó'Fágáin C., Cummins P.M., O'Connor B.F., 2011. Gel-Filtration Chromatography. In: Walls D., Loughran S. (eds) *Protein Chromatography. Methods in Molecular Biology (Methods and Protocols)*, vol 681, Humana Press, Totowa, pp.25-33. https://doi.org/10.1007/978-1-60761-913-0_2

Orozco, B.M., Gladfelter, H.J., Settlage, S.B., Eagle, P.A., Gentry, R.N. and Hanley-Bowdoin, L., 1998. MultipleCisElements Contribute to Geminivirus Origin Function. Virology, 242(2), pp.346-356. https://doi.org/10.1006/viro.1997.9013

Orozco, B.M. and Hanley-Bowdoin, L., 1998. Conserved sequence and structural motifs contribute to the DNA binding and cleavage activities of a geminivirus replication protein. J. Biol. Chem., 273(38), pp.24448-24456. https://doi.org/10.1074/jbc.273.38.24448

Orozco, B.M., Kong, L.J., Batts, L.A., Elledge, S. and Hanley-Bowdoin, L., 2000. The multifunctional character of a geminivirus replication protein is reflected by its complex oligomerization properties. J. Biol. Chem., 275(9), pp.6114-6122. https://doi.org/10.1074/jbc.275.9.6114

Orozco, B.M., Miller, A.B., Settlage, S.B. and Hanley-Bowdoin, L., 1997. Functional domains of a geminivirus replication protein. J. Biol. Chem., 272(15), pp.9840-9846. https://doi.org/10.1074/jbc.272.15.9840

Padidam, M., Beachy, R.N. and Fauquet, C.M., 1995. Tomato leaf curl geminivirus from India has a bipartite genome and coat protein is not essential for infectivity. J. Gen. Virol., 76(1), pp.25-35. https://doi.org/10.1099/0022-1317-76-1-25

Padidam, M., Beachy, R.N. and Fauquet, C.M., 1996. The role of AV2 ("precoat") and coat protein in viral replication and movement in tomato leaf curl geminivirus. Virology, 224(2), pp.390-404. https://doi.org/10.1006/viro.1996.0546

Padidam, M., Sawyer, S. and Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology, 265(2), pp.218-225. https://doi.org/10.1006/viro.1999.0056

Pan, H., Chu, D., Yan, W., Su, Q., Liu, B., Wang, S., Wu, Q., Xie, W., Jiao, X., Li, R. and Yang, N., 2012. Rapid spread of Tomato yellow leaf curl virus in China is aided

differentially by two invasive whiteflies. PloS one, 7(4), p.e34817. https://doi.org/10.1371/journal.pone.0034817

Pan, S.H. and Malcolm, B.A., 2000. Reduced background expression and improved plasmid stability with pET vectors in BL21 (DE3). Biotechniques, 29(6), pp.1234-1238. https://doi.org/10.2144/00296st03

Pantazatos, D., Kim, J.S., Klock, H.E., Stevens, R.C., Wilson, I.A., Lesley, S.A. and Woods, V.L., 2004. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. Proc. Natl. Acad. Sci., 101(3), pp.751-756. https://doi.org/10.1073/pnas.0307204101

Pasumarthy, K.K., Choudhury, N.R. and Mukherjee, S.K., 2010. Tomato leaf curl Kerala virus (ToLCKeV) AC3 protein forms a higher order oligomer and enhances ATPase activity of replication initiator protein (Rep/AC1). Virology J., 7, 128. https://doi.org/10.1186/1743-422X-7-128

Peretz, Y., Eybishtz, A. and Sela, I., 2011. Silencing of ORFs C2 and C4 of tomato yellow leaf curl virus engenders resistant or tolerant plants. Open Virol. J., 5, p.141. https://dx.doi.org/10.2174%2F1874357901105010141

Perilla-Henao, L.M. and Casteel, C.L., 2016. Vector-borne bacterial plant pathogens: interactions with hemipteran insects and plants. Front. Plant Sci., 7, 1163. https://doi.org/10.3389/fpls.2016.01163

Piedra-Aguilera, Á., Jiao, C., Luna, A.P., Villanueva, F., Dabad, M., Esteve-Codina, A., Díaz-Pendón, J.A., Fei, Z., Bejarano, E.R. and Castillo, A.G., 2019. Integrated single-base resolution maps of transcriptome, sRNAome and methylome of Tomato yellow leaf curl virus (TYLCV) in tomato. Sci. Rep., 9, 2863. https://doi.org/10.1038/s41598-019-39239-6

Pietersen, G. and Smith, M.F., 2002. Tomato yellow leaf curl virus resistant tomatoes show resistance to tomato curly stunt virus. Plant Dis., 86(5), pp.528-534. https://doi.org/10.1094/PDIS.2002.86.5.528

Pietersen, G., Idris, A.M., Krüger, K. and Brown, J.K., 2000. Tomato curly stunt virus, a new begomovirus of tomato within the Tomato yellow leaf curl virus-IS cluster in

South Africa. Plant Dis., 84(7), pp.810-810. https://doi.org/10.1094/PDIS.2000.84.7.810B

Pietersen, G., Idris, A.M., Krüger, K. and Brown, J.K., 2008. Characterization of Tomato curly stunt virus: a new tomato-infecting begomovirus from South Africa. Plant Pathol., 57(5), pp.809-818. https://doi.org/10.1111/j.1365-3059.2008.01882.x

Pita, J.S., Fondong, V.N., Sangare, A., Otim-Nape, G.W., Ogwal, S. and Fauquet, C.M., 2001. Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. J. Gen. Virol., 82(3), pp.655-665. https://doi.org/10.1099/0022-1317-82-3-655

Pitaksutheepong, C., Vimoltat, A., Nathwong, B. and Attathom, S., 2007. The N-terminal 62 amino acid residues of the coat protein of Tomato yellow leaf curl Thailand virus are responsible for DNA binding. J. Gen. Plant Pathol., 73(1), pp.72-75. https://doi.org/10.1007/s10327-006-0317-5

Pooggin, M.M., 2013. How can plant DNA viruses evade siRNA-directed DNA methylation and silencing?. Int. J. Mol. Sci., 14(8), pp.15233-15259. https://doi.org/10.3390/ijms140815233

Poornima Priyadarshini, C.G., Ambika, M.V., Tippeswamy, R. and Savithri, H.S., 2011. Functional characterization of coat protein and V2 involved in cell to cell movement of Cotton leaf curl Kokhran virus-Dabawali. PloS one, 6(11), p.e26929. https://doi.org/10.1371/journal.pone.0026929

Qing, G., Ma, L.C., Khorchid, A., Swapna, G.V.T., Mal, T.K., Takayama, M.M., Xia, B., Phadtare, S., Ke, H., Acton, T. and Montelione, G.T., 2004. Cold-shock induced high-yield protein production in Escherichia coli. Nat. Biotechnol., 22(7), pp.877-882. https://doi.org/10.1038/nbt984

Raiola, A., Rigano, M.M., Calafiore, R., Frusciante, L. and Barone, A., 2014. Enhancing the health-promoting effects of tomato fruit for biofortified food. Mediators Inflamm., 2014 https://doi.org/10.1155/2014/139873

Rey, C. and Vanderschuren, H., 2017. Cassava mosaic and brown streak diseases: current perspectives and beyond. Annu. Rev. Virol., 4, pp.429-452. https://doi.org/10.1146/annurev-virology-101416-041913

Rey, M.E., Ndunguru, J., Berrie, L.C., Paximadis, M., Berry, S., Cossa, N., Nuaila, V.N., Mabasa, K.G., Abraham, N., Rybicki, E.P. and Martin, D., 2012. Diversity of dicotyledenous-infecting geminiviruses and their associated DNA molecules in Southern Africa, including the South-West Indian Ocean Islands. Viruses, 4(9), pp.1753-1791. https://doi.org/10.3390/v4091753

Richter, K.S., Ende, L. and Jeske, H., 2015. Rad54 is not essential for any geminiviral replication mode in planta. Plant Mol. Biol., 87(1), pp.193-202. https://doi.org/10.1007/s11103-014-0270-1

Rigden, J.E., Dry, I.B., Mullineaux, P.M. and Rezaian, M.A., 1993. Mutagenesis of the virion-sense open reading frames of tomato leaf curl geminivirus. Virology, 193(2), pp.1001-1005. https://doi.org/10.1006/viro.1993.1215

Riggs, P., 2000. Expression and purification of recombinant proteins by fusion to maltose-binding protein. Mol. Biotechnol., 15(1), pp.51-63. https://doi.org/10.1385/MB:15:1:51

Rishishwar, R. and Dasgupta, I., 2019. Suppressors of RNA silencing encoded by geminiviruses and associated DNA satellites. Virusdisease, 30(1), pp.58-65. https://doi.org/10.1007/s13337-018-0418-8

Rivera-Calzada, A. and Carroni, M., 2019. Technical Advances in Cryo-Electron Microscopy. Front. Mol. Biosci., 6, 72. https://doi.org/10.3389/fmolb.2019.00072

Rizvi, I., Choudhury, N.R. and Tuteja, N., 2015. Insights into the functional characteristics of geminivirus rolling-circle replication initiator protein and its interaction with host factors affecting viral DNA replication. Arch. Virol., 160(2), pp.375-387. https://doi.org/10.1007/s00705-014-2297-7

Rodríguez-Negrete, E.A., Carrillo-Tripp, J. and Rivera-Bustamante, R.F., 2009. RNA silencing against geminivirus: complementary action of posttranscriptional gene silencing and transcriptional gene silencing in host recovery. J. Virol., 83(3), pp.1332-1340. https://doi.org/10.1128/JVI.01474-08

Rodríguez-Negrete, E., Lozano-Durán, R., Piedra-Aguilera, A., Cruzado, L., Bejarano, E.R. and Castillo, A.G., 2013. Geminivirus R ep protein interferes with the plant DNA

methylation machinery and suppresses transcriptional gene silencing. New Phytol., 199(2), pp.464-475. https://doi.org/10.1111/nph.12286

Rojas, M.R., Jiang, H., Salati, R., Xoconostle-Cázares, B., Sudarshana, M.R., Lucas, W.J. and Gilbertson, R.L., 2001. Functional analysis of proteins involved in movement of the monopartite begomovirus, Tomato yellow leaf curl virus. Virology, 291(1), pp.110-125. https://doi.org/10.1006/viro.2001.1194

Rojas, M.R., Macedo, M.A., Maliano, M.R., Soto-Aguilar, M., Souza, J.O., Briddon, R.W., Kenyon, L., Rivera Bustamante, R.F., Zerbini, F.M., Adkins, S. and Legg, J.P., 2018. World management of geminiviruses. Annu. Rev. Phytopathol., 56, pp.637-677. https://doi.org/10.1146/annurev-phyto-080615-100327

Rojas, M.R., Noueiry, A.O., Lucas, W.J. and Gilbertson, R.L., 1998. Bean dwarf mosaic geminivirus movement proteins recognize DNA in a form-and size-specific manner. Cell, 95(1), pp.105-113. https://doi.org/10.1016/S0092-8674(00)81786-9

Rosano, G.L. and Ceccarelli, E.A., 2014. Recombinant protein expression in Escherichia coli: advances and challenges. Front. Microbiol., 5, p.172. https://doi.org/10.3389/fmicb.2014.00172

Roshan, P., Kulshreshtha, A. and Hallan, V., 2017. Identification of host cellular targets of AC4 and AV2 proteins of tomato leaf curl palampur virus and their sub-cellular localization studies. VirusDisease, 28(4), pp.390-400. https://doi.org/10.1007/s13337-017-0405-5

Roshan, P., Kulshreshtha, A., Kumar, S., Purohit, R. and Hallan, V., 2018. AV2 protein of tomato leaf curl Palampur virus promotes systemic necrosis in *Nicotiana benthamiana* and interacts with host Catalase2. Sci. Rep., 8, 1273. https://doi.org/10.1038/s41598-018-19292-3

Roumagnac, P., Lett, J.M., Fiallo-Olivé, E., Navas-Castillo, J., Zerbini, F.M., Martin, D.P. and Varsani, A., 2021. Establishment of five new genera in the family Geminiviridae: Citlodavirus, Maldovirus, Mulcrilevirus, Opunvirus, and Topilevirus. Arch. Virol., 167, pp.695–710. https://doi.org/10.1007/s00705-021-05309-2, pp.1-16

Ruhel, R. and Chakraborty, S., 2019. Multifunctional roles of geminivirus encoded replication initiator protein. VirusDisease, 30(1), pp.66-73. https://doi.org/10.1007/s13337-018-0458-0

Ruhel, R., Mazumder, M., Gnanasekaran, P., Kumar, M., Gourinath, S. and Chakraborty, S., 2021. Functional implications of residues of the B′ motif of geminivirus replication initiator protein in its helicase activity. FEBS J., 288(22), pp.6492-6509. https://doi.org/10.1111/febs.16053

Ruiz-Medrano, R., Guevara-Gonzalez, R.G., Argüello-Astorga, G.R., Monsalve-Fonnegra, Z., Herrera-Estrella, L.R. and Rivera-Bustamante, R.F., 1999. Identification of a sequence element involved in AC2-mediated transactivation of the pepper huasteco virus coat protein gene. Virology, 253(2), pp.162-169. https://doi.org/10.1006/viro.1998.9484

Saeed, F., Sattar, M.N., Hameed, U., Ilyas, M., Haider, M.S., Hamza, M., Mansoor, S. and Amin, I., 2018. Infectivity of okra enation leaf curl virus and the role of its V2 protein in pathogenicity. Virus Res., 255, pp.90-94. https://doi.org/10.1016/j.virusres.2018.07.007

Saini, P., Wani, S.I., Kumar, R., Chhabra, R., Chimni, S.S. and Sareen, D., 2014. Trigger factor assisted folding of the recombinant epoxide hydrolases identified from C. pelagibacter and S. nassauensis. Protein Expr. Purif., 104, pp.71-84. https://doi.org/10.1016/j.pep.2014.09.004

Samuelson, J.C., 2011. Recent developments in difficult protein expression: a guide to E. coli strains, promoters, and relevant host mutations. In: Evans, Jr. T.C., Xu, M.Q. (eds), *Heterologous Gene Expression in E.coli. Methods in Molecular Biology (Methods and Protocols)*, vol 705, Humana Press, Totowa, pp.195-209. https://doi.org/10.1007/978-1-61737-967-3_11

Sánchez-Durán, M.A., Dallas, M.B., Ascencio-Ibañez, J.T., Reyes, M.I., Arroyo-Mateos, M., Ruiz-Albert, J., Hanley-Bowdoin, L. and Bejarano, E.R., 2011. Interaction between geminivirus replication protein and the SUMO-conjugating enzyme is required for viral infection. J. Virol., 85(19), pp.9789-9800. https://doi.org/10.1128/JVI.02566-10

Sande, O.F., Orílio, A.F., Chipiringo, B.A., Xavier, C.A. and Zerbini, F.M., 2021. Speciation driven by recombination in the evolution of tomato curly stunt virus in Mozambique. Plant Pathol., 70(4), pp.994-1002. https://doi.org/10.1111/ppa.13342

Sanders, C.M., Kovalevskiy, O.V., Sizov, D., Lebedev, A.A., Isupov, M.N. and Antson, A.A., 2007. Papillomavirus E1 helicase assembly maintains an asymmetric state in the absence of DNA and nucleotide cofactors. Nucleic Acids Res., 35(19), pp.6451-6457. https://doi.org/10.1093/nar/gkm705

Santosh, V., Musayev, F.N., Jaiswal, R., Zárate-Pérez, F., Vandewinkel, B., Dierckx, C., Endicott, M., Sharifi, K., Dryden, K., Henckaerts, E. and Escalante, C.R., 2020. The Cryo-EM structure of AAV2 Rep68 in complex with ssDNA reveals a malleable AAA+ machine that can switch between oligomeric states. Nucleic Acids Res., 48(22), pp.12983-12999. https://doi.org/10.1093/nar/gkaa1133

Sardo, L., Lucioli, A., Tavazza, M., Masenga, V., Tavazza, R., Accotto, G.P. and Noris, E., 2011. An RGG sequence in the replication-associated protein (Rep) of Tomato yellow leaf curl Sardinia virus is involved in transcriptional repression and severely impacts resistance in Rep-expressing plants. J. Gen. Virol., 92(1), pp.204-209. https://doi.org/10.1099/vir.0.025817-0

Saunders, K., Lucy, A. and Stanley, J., 1991. DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. Nucleic Acids Res., 19(9), pp.2325-2330. https://doi.org/10.1093/nar/19.9.2325

Sawaya, M.R., Guo, S., Tabor, S., Richardson, C.C. and Ellenberger, T., 1999. Crystal structure of the helicase domain from the replicative helicase-primase of bacteriophage T7. Cell, 99(2), pp.167-177. https://doi.org/10.1016/S0092-8674(00)81648-7

Settlage, S.B., Miller, A.B. and Hanley-Bowdoin, L., 1996. Interactions between geminivirus replication proteins. J. Virol., 70(10), pp.6790-6795. https://doi.org/10.1128/jvi.70.10.6790-6795.1996

Settlage, S.B., Miller, A.B., Gruissem, W. and Hanley-Bowdoin, L., 2001. Dual interaction of a geminivirus replication accessory factor with a viral replication protein and a plant cell cycle regulator. Virology, 279(2), pp.570-576. https://doi.org/10.1006/viro.2000.0719

Sharma, P. and Ikegami, M., 2009. Characterization of signals that dictate nuclear/nucleolar and cytoplasmic shuttling of the capsid protein of Tomato leaf curl Java virus associated with DNAβ satellite. Virus Res., 144(1-2), pp.145-153. https://doi.org/10.1016/j.virusres.2009.04.019

Sharma, P. and Ikegami, M., 2010. Tomato leaf curl Java virus V2 protein is a determinant of virulence, hypersensitive response and suppression of posttranscriptional gene silencing. Virology, 396(1), pp.85-93. https://doi.org/10.1016/j.virol.2009.10.012

Sharma, P., Ikegami, M. and Kon, T., 2010. Identification of the virulence factors and suppressors of posttranscriptional gene silencing encoded by Ageratum yellow vein virus, a monopartite begomovirus. Virus Res., 149(1), pp.19-27. https://doi.org/10.1016/j.virusres.2009.12.008

Shen, J., Gai, D., Patrick, A., Greenleaf, W.B. and Chen, X.S., 2005. The roles of the residues on the channel β-hairpin and loop structures of simian virus 40 hexameric helicase. Proc. Natl. Acad. Sci., 102(32), pp.11248-11253. https://doi.org/10.1073/pnas.0409646102

Shivaprasad, P.V., Akbergenov, R., Trinks, D., Rajeswaran, R., Veluthambi, K., Hohn, T. and Pooggin, M.M., 2005. Promoters, transcripts, and regulatory proteins of Mungbean yellow mosaic geminivirus. J. Virol., 79(13), pp.8149-8163. https://doi.org/10.1128/JVI.79.13.8149-8163.2005

Simonne, A.H., do Nascimento Nunes, C. and Brecht, J.K., 2011. 16: Tomato and other solanaceous fruits. In Terry, L.A. (ed), *Health-promoting properties of fruit and vegetables*, CAB International, London, pp.321-324.

Simossis, V.A. and Heringa, J., 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res., 33(suppl_2), pp.W289-W294. https://doi.org/10.1093/nar/gki390

Singh, D.K., Karjee, S., Malik, P.S., Islam, N. and Mukherjee, S.K., 2007. DNA replication and pathogenecity of MYMIV. Communicating current research and educational topics and trends in applied microbiology, 1, pp.155-162.

Singh, D.K., Malik, P.S., Choudhury, N.R. and Mukherjee, S.K., 2008. MYMIV replication initiator protein (Rep): roles at the initiation and elongation steps of MYMIV DNA replication. Virology, 380(1), pp.75-83. https://doi.org/10.1016/j.virol.2008.07.010

Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A., 2007. XtalPred: a web server for prediction of protein crystallizability. Bioinformatics, 23(24), pp.3403-3405. https://doi.org/10.1093/bioinformatics/btm477

Sørensen, H.P. and Mortensen, K.K., 2005. Advanced genetic strategies for recombinant protein expression in Escherichia coli. J. Biotechnol., 115(2), pp.113-128. https://doi.org/10.1016/j.jbiotec.2004.08.004

Stanley, J., Latham, J.R., Pinner, M.S., Bedford, I. and Markham, P.G., 1992. Mutational analysis of the monopartite geminivirus beet curly top virus. Virology, 191(1), pp.396-405. https://doi.org/10.1016/0042-6822(92)90201-y

Stanton P., 2004. Gel Filtration Chromatography. In: Aguilar M.-I. (ed) *HPLC of Peptides and Proteins. Methods in Molecular Biology*, vol 251, Springer, Totowa, pp. 55-74. https://doi.org/10.1385/1-59259-742-4:55

Stern-Ginossar, N., 2015. Decoding viral infection by ribosome profiling. J. Virol., 89(12), pp.6164-6166. https://doi.org/10.1128/JVI.02528-14

Studier, F.W. and Moffatt, B.A., 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol., 189(1), pp.113-130. https://doi.org/10.1016/0022-2836(86)90385-2

Sundaresan, G., Das, S.S., Tripathi, A., Mukherjee, S.K. and Sanan-Mishra, N., 2020. Evaluating the strength of RNA silencing suppressor proteins encoded by two geminiviruses using assay based on reversal of GFP silencing. Australas. Plant Pathol., 49(2), pp.95-106. https://doi.org/10.1007/s13313-019-00678-4

Sung, Y.K. and Coutts, R.H., 1996. Potato yellow mosaic geminivirus AC2 protein is a sequence non-specific DNA binding protein. FEBS Lett., 383(1-2), pp.51-54. https://doi.org/10.1016/0014-5793(96)00217-7

Sunter, G. and Bisaro, D.M., 1997. Regulation of a geminivirus coat protein promoter by AL2 protein (TrAP): evidence for activation and derepression mechanisms. Virology, 232(2), pp.269-280. https://doi.org/10.1006/viro.1997.8549

Sunter, G., Hartitz, M.D. and Bisaro, D.M., 1993. Tomato golden mosaic virus leftward gene expression: autoregulation of geminivirus replication protein. Virology, 195(1), pp.275-280. https://doi.org/10.1006/viro.1993.1374

Tan, P.H., Wong, S.M., Wu, M., Bedford, I.D., Saunders, K. and Stanley, J., 1995. Genome organization of ageratum yellow vein virus, a monopartite whitefly-transmitted geminivirus isolated from a common weed. J. Gen. Virol., 76(12), pp.2915-2922. https://doi.org/10.1099/0022-1317-76-12-2915

Tarasova, E. and Khayat, R., 2022. A Structural Perspective of Reps from CRESS-DNA Viruses and Their Bacterial Plasmid Homologues. Viruses, 14(1), p.37. https://doi.org/10.3390/v14010037

Tarasova, E., Dhindwal, S., Popp, M., Hussain, S. and Khayat, R., 2021. Mechanism of DNA Interaction and Translocation by the Replicase of a Circular Rep-Encoding Single-Stranded DNA Virus. Mbio, 12(4), pp.e00763-21. https://doi.org/10.1128/mBio.00763-21

Teixeira, R.M., Ferreira, M.A., Raimundo, G.A. and Fontes, E.P., 2021. Geminiviral triggers and suppressors of plant antiviral immunity. Microorganisms, 9(4), p.775. https://doi.org/10.3390/microorganisms9040775

Teng, K., Chen, H., Lai, J., Zhang, Z., Fang, Y., Xia, R., Zhou, X., Guo, H. and Xie, Q., 2010. Involvement of C4 protein of beet severe curly top virus (family Geminiviridae) in virus movement. PloS one, 5(6), p.e11280. https://doi.org/10.1371/journal.pone.0011280

Thresh, J.M., Otim-Nape, G.W., Legg, J.P. and Fargette, D., 1997. African cassava mosaic virus disease: the magnitude of the problem. Afr. J. Root Tuber Crops, 2(1/2), pp.13-19.

Tompkins, K.J., Houtti, M., Litzau, L.A., Aird, E.J., Everett, B.A., Nelson, A.T., Pornschloegl, L., Limón-Swanson, L.K., Evans III, R.L., Evans, K. and Shi, K., 2021. Molecular underpinnings of ssDNA specificity by Rep HUH-endonucleases and

implications for HUH-tag multiplexing and engineering. Nucleic Acids Res., 49(2), pp.1046-1064. https://dx.doi.org/10.1093%2Fnar%2Fgkaa1248

Torres-Pacheco, I., Garzón-Tiznado, J.A., Herrera-Estrella, L. and Rivera-Bustamante, R.F., 1993. Complete nucleotide sequence of pepper huasteco virus: analysis and comparison with bipartite geminiviruses. J. Gen. Virol., 74(10), pp.2225-2231. https://doi.org/10.1099/0022-1317-74-10-2225

Trinks, D., Rajeswaran, R., Shivaprasad, P.V., Akbergenov, R., Oakeley, E.J., Veluthambi, K., Hohn, T. and Pooggin, M.M., 2005. Suppression of RNA silencing by a geminivirus nuclear protein, AC2, correlates with transactivation of host genes. J. Virol., 79(4), pp.2517-2527. https://doi.org/10.1128/JVI.79.4.2517-2527.2005

Unseld, S., Höhnle, M., Ringel, M. and Frischmuth, T., 2001. Subcellular targeting of the coat protein of African cassava mosaic geminivirus. Virology, 286(2), pp.373-383. https://doi.org/10.1006/viro.2001.1003

Vadivukarasi, T., Girish, K.R. and Usha, R., 2007. Sequence and recombination analyses of the geminivirus replication initiator protein. J. Biosci., 32(1), pp.17-29. https://doi.org/10.1007/s12038-007-0003-6

Vanitharani, R., Chellappan, P., Pita, J.S. and Fauquet, C.M., 2004. Differential roles of AC2 and AC4 of cassava geminiviruses in mediating synergism and suppression of posttranscriptional gene silencing. J. Virol., 78(17), pp.9487-9498. https://doi.org/10.1128/JVI.78.17.9487-9498.2004

Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P. and Martin, D.P., 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. Virol. J., 6(1), pp.1-12. https://doi.org/10.1186/1743-422X-6-36

Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S. and Windram, O.P., 2008. Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. The J. Gen. Virol., 89(Pt 9), pp.2063-2074. https://dx.doi.org/10.1099%2Fvir.0.2008%2F003590-0

Vega-Rocha, S., Byeon, I.J.L., Gronenborn, B., Gronenborn, A.M. and Campos-Olivas, R., 2007a. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J. Mol. Biol., 367(2), pp.473-487. https://doi.org/10.1016/j.jmb.2007.01.002

Vega-Rocha, S., Gronenborn, B., Gronenborn, A.M. and Campos-Olivas, R., 2007b. Solution structure of the endonuclease domain from the master replication initiator protein of the nanovirus faba bean necrotic yellows virus and comparison with the corresponding geminivirus and circovirus structures. Biochemistry, 46(21), pp.6201-6212. https://doi.org/10.1021/bi700159q

Velten, J., Morey, K.J. and Cazzonelli, C.I., 2005. Plant viral intergenic DNA sequence repeats with transcription enhancing activity. Virol. J., 2(1), pp.1-10. https://doi.org/10.1186/1743-422X-2-16

Venkataraman, S. and Selvarajan, R., 2019. Recent advances in understanding the replication initiator protein of the ssDNA plant viruses of the family Nanoviridae. VirusDisease, 30(1), pp.22-31. https://doi.org/10.1007/s13337-019-00514-9

Wang, B., Li, F., Huang, C., Yang, X., Qian, Y., Xie, Y. and Zhou, X., 2014b. V2 of tomato yellow leaf curl virus can suppress methylation-mediated transcriptional gene silencing in plants. J. Gen. Virol., 95(1), pp.225-230. https://doi.org/10.1099/vir.0.055798-0

Wang, B., Yang, X., Wang, Y., Xie, Y. and Zhou, X., 2018. Tomato yellow leaf curl virus V2 interacts with host histone deacetylase 6 to suppress methylation-mediated transcriptional gene silencing in plants. J. Virol., 92(18), pp.e00036-18. https://doi.org/10.1128/jvi.00036-18

Wang, L., Ding, Y., He, L., Zhang, G., Zhu, J.K. and Lozano-Duran, R., 2020. A virus-encoded protein suppresses methylation of the viral genome through its interaction with AGO4 in the Cajal body. Elife, 9, p.e55542. https://doi.org/10.7554/elife.55542

Wang, W.C., Hsu, Y.H., Lin, N.S., Wu, C.Y., Lai, Y.C. and Hu, C.C., 2013. A novel prokaryotic promoter identified in the genome of some monopartite begomoviruses. PloS one, 8(7), p.e70037. https://doi.org/10.1371/journal.pone.0070037

Wang, W.C., Wu, C.Y., Lai, Y.C., Lin, N.S., Hsu, Y.H. and Hu, C.C., 2014a. Characterization of the cryptic AV3 promoter of Ageratum yellow vein virus in prokaryotic and eukaryotic systems. Plos one, 9(9), p.e108608. https://doi.org/10.1371/journal.pone.0108608

Wang, Y., Wu, Y., Gong, Q., Ismayil, A., Yuan, Y., Lian, B., Jia, Q., Han, M., Deng, H., Hong, Y. and Hanley-Bowdoin, L., 2019. Geminiviral V2 protein suppresses transcriptional gene silencing through interaction with AGO4. J. Virol., 93(6), pp.e01675-18. https://dx.doi.org/10.1128%2FJVI.01675-18

Ward, B.M., Medville, R., Lazarowitz, S.G. and Turgeon, R., 1997. The geminivirus BL1 movement protein is associated with endoplasmic reticulum-derived tubules in developing phloem cells. J. Virol., 71(5), pp.3726-3733. https://doi.org/10.1128/jvi.71.5.3726-3733.1997

Wartig, L., Kheyr-Pour, A., Noris, E., De Kouchkovsky, F., Jouanneau, F., Gronenborn, B. and Jupin, I., 1997. Genetic analysis of the monopartite tomato yellow leaf curl geminivirus: roles of V1, V2, and C2 ORFs in viral pathogenesis. Virology, 228(2), pp.132-140. https://doi.org/10.1006/viro.1996.8406

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L. and Lepore, R., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res., 46(W1), pp.W296-W303. https://doi.org/10.1093/nar/gky427

Whelan, F., Stead, J.A., Shkumatov, A.V., Svergun, D.I., Sanders, C.M. and Antson, A.A., 2012. A flexible brace maintains the assembly of a hexameric replicative helicase during DNA unwinding. Nucleic Acids Res., 40(5), pp.2271-2283. https://doi.org/10.1093/nar/gkr906

Wingfield, P., 1998. Protein Precipitation Using Ammonium Sulfate. Curr. Protoc. Protein Sci., 13: A.3F.1-A.3F.8. https://doi.org/10.1002/0471140864.psa03fs13

Xu, X., Qian, Y., Wang, Y., Li, Z. and Zhou, X., 2019. Iterons homologous to helper geminiviruses are essential for efficient replication of betasatellites. J. Virol., 93(5), pp.e01532-18. https://dx.doi.org/10.1128%2FJVI.01532-18

Yan, Z., Wolters, A.M.A., Navas-Castillo, J. and Bai, Y., 2021. The global dimension of tomato yellow leaf curl disease: Current status and breeding perspectives. Microorganisms, 9(4), p.740. https://doi.org/10.3390/microorganisms9040740

Yang, S.J., Carter, S.A., Cole, A.B., Cheng, N.H. and Nelson, R.S., 2004. A natural variant of a host RNA-dependent RNA polymerase is associated with increased susceptibility to viruses by Nicotiana benthamiana. Proc. Natl. Acad. Sci., 101(16), pp.6297-6302. https://doi.org/10.1073/pnas.0304346101

Yang, Y., Liu, T., Shen, D., Wang, J., Ling, X., Hu, Z., Chen, T., Hu, J., Huang, J., Yu, W. and Dou, D., 2019. Tomato yellow leaf curl virus intergenic siRNAs target a host long noncoding RNA to modulate disease symptoms. PLoS Pathog., 15(1), p.e1007534. https://doi.org/10.1371/journal.ppat.1007534

Yang, X., Ren, Y., Sun, S., Wang, D., Zhang, F., Li, D., Li, S. and Zhou, X., 2018. Identification of the potential virulence factors and RNA silencing suppressors of mulberry mosaic dwarf-associated geminivirus. Viruses, 10(9), 472. https://doi.org/10.3390/v10090472

Yang, X., Wang, Y., Guo, W., Xie, Y., Xie, Q., Fan, L. and Zhou, X., 2011. Characterization of small interfering RNAs derived from the geminivirus/betasatellite complex using deep sequencing. PLoS One, 6(2), p.e16928. https://doi.org/10.1371/journal.pone.0016928

Yoon-Robarts, M., Blouin, A.G., Bleker, S., Kleinschmidt, J.A., Aggarwal, A.K., Escalante, C.R. and Linden, R.M., 2004. Residues within the B′ motif are critical for DNA binding by the superfamily 3 helicase Rep40 of adeno-associated virus type 2. J. Biol. Chem., 279(48), pp.50472-50481. https://doi.org/10.1074/jbc.M403900200

Yukawa, A., Iino, R., Watanabe, R., Hayashi, S. and Noji, H., 2015. Key chemical factors of arginine finger catalysis of F1-ATPase clarified by an unnatural amino acid mutation. Biochemistry, 54(2), pp.472-480. https://doi.org/10.1021/bi501138b

Zerbini, F.M., Briddon, R.W., Idris, A., Martin, D.P., Moriones, E., Navas-Castillo, J., Rivera-Bustamante, R., Roumagnac, P., Varsani, A. and Consortium, I.R., 2017. ICTV virus taxonomy profile: Geminiviridae. J. Gen. Virol., 98(2), p.131. https://dx.doi.org/10.1099%2Fjgv.0.000738

Zhan, X., Haley, A., Richardson, K. and Morris, B., 1991. Analysis of the potential promoter sequences of African cassava mosaic virus by transient expression of the β-glucuronidase gene. J. Gen. Virol., 72(11), pp.2849-2852. https://doi.org/10.1099/0022-1317-72-11-2849

Zhan, B., Zhao, W., Li, S., Yang, X. and Zhou, X., 2018. Functional scanning of apple geminivirus proteins as symptom determinants and suppressors of posttranscriptional gene silencing. Viruses, 10(9), 488. https://doi.org/10.3390/v10090488

Zhang, J., Dong, J., Xu, Y. and Wu, J., 2012. V2 protein encoded by Tomato yellow leaf curl China virus is an RNA silencing suppressor. Virus Res., 163(1), pp.51-58. https://doi.org/10.1016/j.virusres.2011.08.009

Zhao, W., Ji, Y., Wu, S., Ma, X., Li, S., Sun, F., Cheng, Z., Zhou, Y. and Fan, Y., 2018. Single amino acid in V2 encoded by TYLCV is responsible for its self-interaction, aggregates and pathogenicity. Sci. Rep., 8(1), pp.1-11. https://doi.org/10.1038/s41598-018-21446-2

Zhao, X., Li, G. and Liang, S., 2013. Several affinity tags commonly used in chromatographic purification. J. Anal. Methods Chem., 2013:581093. https://doi.org/10.1155/2013/581093

Zhao, W., Wu, S., Barton, E., Fan, Y., Ji, Y., Wang, X. and Zhou, Y., 2020. Tomato yellow leaf curl virus V2 protein plays a critical role in the nuclear export of V1 protein and viral systemic infection. Front. Microbiol., 11, 1243. https://doi.org/10.3389/fmicb.2020.01243

Zheng, L., Baumann, U. and Reymond, J.L., 2004. An efficient one-step site-directed and site-saturation mutagenesis protocol. Nucleic Acids Res., 32(14), pp.e115-e115. https://doi.org/10.1093/nar/gnh110

Zinga, I., Chiroleu, F., Legg, J., Lefeuvre, P., Komba, E.K., Semballa, S., Yandia, S.P., Mandakombo, N.B., Reynaud, B. and Lett, J.M., 2013. Epidemiological assessment of cassava mosaic disease in Central African Republic reveals the importance of mixed viral infection and poor health of plant cuttings. Crop Prot., 44, pp.6-12. https://doi.org/10.1016/j.cropro.2012.10.010

Zrachya, A., Glick, E., Levy, Y., Arazi, T., Citovsky, V. and Gafni, Y., 2007. Suppressor of RNA silencing encoded by Tomato yellow leaf curl virus-Israel. Virology, 358(1), pp.159-165. https://doi.org/10.1016/j.virol.2006.08.016

Appendix (A) 1

ToCSV Infectious clone construct nt sequences

>pCAMBIA2300_V30

catgccaaccacagggttcccctcgggatcaaagtactttgatccaacccctccgctgctatagtgca
gtcggcttctgacgttcagtgcagccgtcttctgaaaacgacatgtcgcacaagtcctaagttacgcg
acaggctgccgccctgccctttttcctggcgttttcttgtcgcgtgttttagtcgcataaagtagaata
cttgcgactagaaccggagacattacgccatgaacaagagcgccgccgctggcctgctgggctatgcc
cgcgtcagcaccgacgaccaggacttgaccaaccaacgggccgaactgcacgcggccggctgcaccaa
gctgttttccgagaagatcaccggcaccaggcgcgaccgcccggagctggccaggatgcttgaccacc
tacgccctggcgacgttgtgacagtgaccaggctagaccgcctggcccgcagcacccgcgacctactg
gacattgccgagcgcatccaggaggccggcgcgggcctgcgtagcctggcagagccgtgggccgacac
caccacgccggccggccgcatggtgttgaccgtgttcgccggcattgccgagttcgagcgttccctaa
tcatcgaccgcacccggagcgggcgcgaggccgccaaggcccgaggcgtgaagtttggcccccgccct
accctcaccccggcacagatcgcgcacgcccgcgagctgatcgaccaggaaggccgcaccgtgaaaga
ggcggctgcactgcttggcgtgcatcgctcgaccctgtaccgcgcacttgagcgcagcgaggaagtga
cgcccaccgaggccaggcggcgcggtgccttccgtgaggacgcattgaccgaggccgacgccctggcg
gccgccgagaatgaacgccaagaggaacaagcatgaaaccgcaccaggacggccaggacgaaccgttt
ttcattaccgaagagatcgaggcggagatgatcgcggccgggtacgtgttcgagccgcccgcgcacgt
ctcaaccgtgcggctgcatgaaatcctggccggtttgtctgatgccaagctggcggcctggccggcca
gcttggccgctgaagaaaccgagcgccgccgtctaaaaaggtgatgtgtatttgagtaaaacagcttg
cgtcatgcggtcgctgcgtatatgatgcgatgagtaaataaacaaatacgcaaggggaacgcatgaag
gttatcgctgtacttaaccagaaaggcgggtcaggcaagacgaccatcgcaacccatctagcccgcgc
cctgcaactcgccggggccgatgttctgttagtcgattccgatccccagggcagtgcccgcgattggg
cggccgtgcgggaagatcaaccgctaaccgttgtcggcatcgaccgcccgacgattgaccgcgacgtg
aaggccatcggccggcgcgacttcgtagtgatcgacggagcgccccaggcggcggacttggctgtgtc
cgcgatcaaggcagccgacttcgtgctgattccggtgcagccaagcccttacgacatatgggccaccg
ccgacctggtggagctggttaagcagcgcattgaggtcacggatggaaggctacaagcggcctttgtc
gtgtcgcgggcgatcaaaggcacgcgcatcggcggtgaggttgccgaggcgctggccgggtacgagct
gcccattcttgagtcccgtatcacgcagcgcgtgagctacccaggcactgccgccgccggcacaaccg
ttcttgaatcagaacccgagggcgacgctgcccgcgaggtccaggcgctggccgctgaaattaaatca
aaactcatttgagttaatgaggtaaagagaaaatgagcaaaagcacaaacacgctaagtgccggccgt
ccgagcgcacgcagcagcaaggctgcaacgttggccagcctggcagacacgccagccatgaagcgggt
caactttcagttgccggcggaggatcacaccaagctgaagatgtacgcggtacgccaaggcaagacca
ttaccgagctgctatctgaatacatcgcgcagctaccagagtaaatgagcaaatgaataaatgagtag
atgaattttagcggctaaaggaggcggcatggaaaatcaagaacaaccaggcaccgacgccgtggaat
gccccatgtgtggaggaacgggcggttggccaggcgtaagcggctgggttgtctgccggccctgcaat
ggcactggaaccccaagcccgaggaatcggcgtgacggtcgcaaaccatccggcccggtacaaatcg
gcgcggcgctgggtgatgacctggtggagaagttgaaggccgcgcaggccgcccagcggcaacgcatc
gaggcagaagcacgccccggtgaatcgtggcaagcggccgctgatcgaatccgcaaagaatcccggca
accgccggcagccggtgcgccgtcgattaggaagccgcccaagggcgacgagcaaccagatttttcg
ttccgatgctctatgacgtgggcacccgcgatagtcgcagcatcatggacgtggccgttttccgtctg
tcgaagcgtgaccgacgagctggcgaggtgatccgctacgagcttccagacgggcacgtagaggtttc
cgcagggccggccggcatggccagtgtgtgggattacgacctggtactgatggcggtttcccatctaa
ccgaatccatgaaccgataccgggaagggaagggagacaagcccggccgcgtgttccgtccacacgtt
gcggacgtactcaagttctgccggcgagccgatggcggaaagcagaaagacgacctggtagaaacctg
cattcggttaaacaccacgcacgttgccatgcagcgtacgaagaaggccaagaacggccgcctggtga
cggtatccgagggtgaagccttgattagccgctacaagatcgtaaagagcgaaaccgggcggccggag
tacatcgagatcgagctagctgattggatgtaccgcgagatcacagaaggcaagaacccggacgtgct
gacggttcaccccgattacttttgatcgatcccggcatcggccgttttctctaccgcctggcacgcc

gcgccgcaggcaaggcagaagccagatggttgttcaagacgatctacgaacgcagtggcagcgccgga
gagttcaagaagttctgtttcaccgtgcgcaagctgatcgggtcaaatgacctgccggagtacgattt
gaaggaggaggcggggcaggctggcccgatcctagtcatgcgctaccgcaacctgatcgagggcgaag
catccgccggttcctaatgtacggagcagatgctagggcaaattgccctagcaggggaaaaaggtcga
aaaggtctctttcctgtggatagcacgtacattgggaacccaaagccgtacattgggaaccggaaccc
gtacattgggaacccaaagccgtacattgggaaccggtcacacatgtaagtgactgatataaaagaga
aaaaaggcgatttttccgcctaaaactctttaaaacttattaaaactcttaaaacccgcctggcctgt
gcataactgtctggccagcgcacagccgaagagctgcaaaaagcgcctacccttcggtcgctgcgctc
cctacgccccgccgcttcgcgtcggcctatcgcggccgctggccgctcaaaaatggctggcctacggc
caggcaatctaccagggcgcggacaagccgcgccgtcgccactcgaccgccggcgcccacatcaaggc
accctgcctcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtca
cagcttgtctgtaagcggatgccgggagcagacaagcccgtcagggcgcgtcagcgggtgttggcggg
tgtcggggcgcagccatgacccagtcacgtagcgatagcggagtgtatactggcttaactatgcggca
tcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaa
ataccgcatcaggcgctcttccgcttcctcgctcactgactcgctgcgctcggtcgttcggctgcggc
gagcggtatcagctcactcaaaggcggtaatacggttatccacagaatcaggggataacgcaggaaag
aacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgcgttgctggcgttttttcca
taggctccgccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcgaaacccgacag
gactataaagataccaggcgtttccccctggaagctccctcgtgcgctctcctgttccgaccctgccg
cttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagctcacgctgtag
gtatctcagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaaccccccgttcagcccg
accgctgcgccttatccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactg
gcagcagccactggtaacaggattagcagagcgaggtatgtaggcggtgctacagagttcttgaagtg
gtggcctaactacggctacactagaaggacagtatttggtatctgcgctctgctgaagccagttacct
tcggaaaaagagttggtagctcttgatccggcaaacaaaccaccgctggtagcggtggtttttttgtt
tgcaagcagcagattacgcgcagaaaaaaaggatctcaagaagatcctttgatcttttctacggggtc
tgacgctcagtggaacgaaaactcacgttaagggattttggtcatgcattctaggtactaaaacaatt
catccagtaaaatataatattttattttctcccaatcaggcttgatccccagtaagtcaaaaaatagc
tcgacatactgttcttccccgatatcctccctgatcgaccggacgcagaaggcaatgtcataccactt
gtccgccctgccgcttctcccaagatcaataaagccacttactttgccatctttcacaaagatgttgc
tgtctcccaggtcgccgtgggaaaagacaagttcctcttcgggcttttccgtctttaaaaaatcatac
agctcgcgcggatctttaaatggagtgtcttcttcccagttttcgcaatccacatcggccagatcgtt
attcagtaagtaatccaattcggctaagcggctgtctaagctattcgtatagggacaatccgatatgt
cgatggagtgaaagagcctgatgcactccgcatacagctcgataatcttttcagggctttgttcatct
tcatactcttccgagcaaaggacgccatcggcctcactcatgagcagattgctccagccatcatgccg
ttcaaagtgcaggacctttggaacaggcagctttccttccagccatagcatcatgtccttttcccgtt
ccacatcataggtggtccctttataccggctgtccgtcatttttaaatataggttttcattttctccc
accagcttatataccttagcaggagacattccttccgtatctttacgcagcggtatttttcgatcag
ttttttcaattccggtgatattctcattttagccatttattatttccttcctctttttctacagtattt
aaagataccccaagaagctaattataacaagacgaactccaattcactgttccttgcattctaaaacc
ttaaataccagaaaacagctttttcaaagttgttttcaaagttggcgtataacatagtatcgacggag
ccgattttgaaaccgcggtgatcacaggcagcaacgctctgtcatcgttacaatcaacatgctaccct
ccgcgagatcatccgtgtttcaaacccggcagcttagttgccgttcttccgaatagcatcggtaacat
gagcaaagtctgccgccttacaacggctctcccgctgacgccgtcccggactgatgggctgcctgtat
cgagtggtgattttgtgccgagctgccggtcggggagctgttggctggctggtggcaggatatattgt
ggtgtaaacaaattgacgcttagacaacttaataacacattgcggacgtttttaatgtactgaattaa
cgccgaattaattcgggggatctggattttagtactggattttggttttaggaattagaaattttatt
gatagaagtattttacaaatacaaatacatactaagggtttcttatatgctcaacacatgagcgaaac
cctataggaaccctaattcccttatctgggaactactcacacattattatggagaaactcgagcttgt
cgatcgactctagctagaggatcgatccgaaccccagagtcccgctcagaagaactcgtcaagaaggc
gatagaaggcgatgcgctgcgaatcgggagcggcgataccgtaaagcacgaggaagcggtcagcccat

tcgccgccaagctcttcagcaatatcacgggtagccaacgctatgtcctgatagcggtccgccacacc
cagccggccacagtcgatgaatccagaaaagcggccattttccaccatgatattcggcaagcaggcat
cgccatgtgtcacgacgagatcctcgccgtcgggcatgcgcgccttgagcctggcgaacagttcggct
ggcgcgagcccctgatgctcttcgtccagatcatcctgatcgacaagaccggcttccatccgagtacg
tgctcgctcgatgcgatgtttcgcttggtggtcgaatgggcaggtagccggatcaagcgtatgcagcc
gccgcattgcatcagccatgatggatactttctcggcaggagcaaggtgagatgacaggagatcctgc
cccggcacttcgcccaatagcagccagtcccttcccgcttcagtgacaacgtcgagcacagctgcgca
aggaacgcccgtcgtggccagccacgatagccgcgctgcctcgtcctggagttcattcagggcaccgg
acaggtcggtcttgacaaaaagaaccgggcgcccctgcgctgacagccggaacacggcggcatcagag
cagccgattgtctgttgtgcccagtcatagccgaatagcctctccacccaagcggccggagaacctgc
gtgcaatccatcttgttcaatccccatggtcgatcgacagatctgcgaaagctcgagagagatagatt
tgtagagagagactggtgatttcagcgtgtcctctccaaatgaaatgaacttccttatatagaggaag
gtcttgcgaaggatagtgggattgtgcgtcatcccttacgtcagtggagatatcacatcaatccactt
gctttgaagacgtggttggaacgtcttcttttttccacgatgctcctcgtgggtgggggtccatctttg
ggaccactgtcggcagaggcatcttgaacgatagcctttccttttatcgcaatgatggcatttgtaggt
gccaccttccttttctactgtccttttgatgaagtgacagatagctgggcaatggaatccgaggaggt
ttcccgatattacccttttgttgaaaagtctcaatagccctttggtcttctgagactgtatctttgata
ttcttggagtagacgagagtgtcgtgctccaccatgttatcacatcaatccacttgctttgaagacgt
ggttggaacgtcttcttttttccacgatgctcctcgtgggtgggggtccatctttgggaccactgtcgg
cagaggcatcttgaacgatagcctttccttttatcgcaatgatggcatttgtaggtgccaccttccttt
tctactgtccttttgatgaagtgacagatagctgggcaatggaatccgaggaggtttcccgatattac
cctttgttgaaaagtctcaatagccctttggtcttctgagactgtatctttgatattcttggagtaga
cgagagtgtcgtgctccaccatgttggcaagctgctctagccaatacgcaaaccgcctctccccgcgc
gttggccgattcattaatgcagctggcacgacaggtttcccgactggaaagcgggcagtgagcgcaac
gcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtat
gttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatgacatgattacgaattcga
gctccaaagcggccatccgattaatattaccggatggccgtgcccccaaaaaaaactagtccccacgc
actattttatgtcgaccaatcagaatgtcgccttatcgcttagttatctgctattttgtctttatatg
ggaaacttcgcggaagttaccattgtcaagatgtgggatccactgttaaacgaattcccagactccgt
ccacggttttcgttgtatgcttgctgttaaatatctgcagtccgttgaagccacatacgagcccaaca
cattgggccacgatcttatacgggatctgattctggttattagggcaaaagattatgtcgaagcgtcc
cgccgatataatcatttccacgcccgcctcgaaggtgcggagaaggctgaacttcgacagcccgttca
ccagccgtgctgttgtccccattgccccaggcacaaacaagcgacgatcatggacgtacaggcccatg
tatcgaaagcccaagatgtacagaatgtatagaagccctgatgtccctcgaggttgtgaaggcccatg
taagatccagtcgtatgaacagcgcgatgatgtgaagcacaccggtatagttaggtgtatcagtgaca
ttacaaagggtcctggtctgacccatcgcacgggtaaaaggttcgtggtgaagtctgtctatatactc
ggtaaggtatggatggatgagaacatcaagaagcagaaccacactaacaatgttatgtttttcctagt
ccgagatagaaggccctatggaacaagcccaatggactttgggcaggtgtttaacatgtttgataacg
agcccagtacagccacggtgaagaacgatcttcgtgatcgattccaggtgttgcgcaagtttacagcc
accgttgttggtggtccctctggtcttaaggagcaggctcttgtgaagcgtttcttcaggttaaacag
ccatgtaacctataaccatcaggaagcagccaagtacgagaaccataccgagaatgcgttgttgttgt
atatggcatgtactcatgcctctaatcctgtgtatgccacattaaaaatacgtgtatatttttatgac
tcagtgatgaattaataaatattaaattttatttcatgtgtctcaacaacattgatagtgtttacaag
tacattataaagtacatgatcaactgctctaattacattgttaattgaaattacacctaatacatcta
aatacttaaaaacttgaaatttaaatacgcttaagaaacgaccagtctgaggctgtaaggtcgtccag
acttggaagttgagaaaacatttgtgaatccccagttccttccggaggttgtggttgaaccgtatctg
gatcgatatgatgtcgtggttcgtgtggaacggtctgctgtcgtggttgataatcttgaaatagaggg
gatttgttatctcccaaatatagacgccattctgggcctgatgagcagtgatgagttccccggtgcgt
aaatccatggttgatgcagtcgatgtggaggtaatatgtacacccgcaatcaagatcaactcgcctac
gcctgactactctcttcttggctattcgatgctggaccttgattgggacttgagtacagtggctggtt
gagggtgatgaagaccgcattctttatagtccaagccttcagggatgtattcttatcctcatcgagat

attctttatatgaggatgttgggcctggattgcagaggaagatagtgggaatgccgcctttaatttga
attggtttcccgtacttcgtgttgctttgccagtccctttgggccccaaagaattccttccagtgctt
taggtaatgcggatctacgtcatcaatgacgttataccaagcgtcattagaatatacctttggactta
agtctaaatgaccacataagtaattatgtggccctaatgatctagcccaaatagttttgccggaacga
ctatcgccctctattacaatactattaggtctccaaggccgcgcagcggcaccacacacattctcaga
gacccactcttcaagttcctcagggacttgatcaaaagaagaagaagaaaagggagaaatataaggct
ctgggggagcctggaaaatcctatctaaattggtctttaaattatgatactgaaaaataaaatcttta
gggagtttttccctaataatagccatagcggcttcagcggaatctgcgtttaaagcgtcggcgcatgc
gtcgttagcattttggcagcctcctctagcacttctgccgtcgacctggaattctccccattcaagtg
tgtctccgtccttgtcgatgtaggacttgacgtcggagctcgatttagctccctgaatgtttggatgg
aaatgtgctgacctgtttggggataccaggtcgaagaatcggtgattcttgcagttgaattttccctc
gaactgaataagcacgtggagatgaggttccccattttcgtgtagctctctgcagattttgatgtatt
ttttattggtgggtgtttcaaggtttaataattgagagagtgactcttcctttgtgagagagcatttt
ggatatgtgacgaaatagtttttggaatttatttgaaagcgcttcggaggagccatttggtcaatggg
taccgattgactctcaaattgcattctcctggtatatcggtacccaatatatagtgggtaccgaatgg
cacaattgtaatttgtggcaaagtaattggcatttcaaaattcaaactcctaaagcggccatccgatt
aatattaccggatggccgtgcccccaaaaaaaactagtaaaaaagcttggatcctctagagtcgacct
gcaggcatgcaagcttggcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttacc
caacttaatcgccttgcagcacatccccctttcgccagctggcgtaatagcgaagaggcccgcaccga
tcgcccttcccaacagttgcgcagcctgaatggcgaatgctagagcagcttgagcttggatcagattg
tcgtttcccgccttcagtttaaactatcagtgtttgacaggatatattggcgggtaaacctaagagaa
aagagcgtttattagaataacggatatttaaaagggcgtgaaaaggtttatccgttcgtccatttgta
tgtg

>pCAMBIA2300_V22

catgccaaccacagggttcccctcgggatcaaagtactttgatccaacccctccgctgctatagtgca
gtcggcttctgacgttcagtgcagccgtcttctgaaaacgacatgtcgcacaagtcctaagttacgcg
acaggctgccgccctgcccttttcctggcgttttcttgtcgcgtgttttagtcgcataaagtagaata
cttgcgactagaaccggagacattacgccatgaacaagagcgccgccgctggcctgctgggctatgcc
cgcgtcagcaccgacgaccaggacttgaccaaccaacgggccgaactgcacgcggccggctgcaccaa
gctgttttccgagaagatcaccggcaccaggcgcgaccgcccggagctggccaggatgcttgaccacc
tacgccctggcgacgttgtgacagtgaccaggctagaccgcctggcccgcagcacccgcgacctactg
gacattgccgagcgcatccaggaggccggcgcgggcctgcgtagcctggcagagccgtgggccgacac
caccacgccggccggccgcatggtgttgaccgtgttcgccggcattgccgagttcgagcgttccctaa
tcatcgaccgcacccggagcgggcgcgaggccgccaaggcccgaggcgtgaagtttggcccccgccct
accctcaccccggcacagatcgcgcacgcccgcgagctgatcgaccaggaaggccgcaccgtgaaaga
ggcggctgcactgcttggcgtgcatcgctcgaccctgtaccgcgcacttgagcgcagcgaggaagtga
cgcccaccgaggccaggcggcgcggtgccttccgtgaggacgcattgaccgaggccgacgccctggcg
gccgccgagaatgaacgccaagaggaacaagcatgaaaccgcaccaggacggccaggacgaaccgttt
ttcattaccgaagagatcgaggcggagatgatcgcggccgggtacgtgttcgagccgcccgcgcacgt
ctcaaccgtgcggctgcatgaaatcctggccggtttgtctgatgccaagctggcggcctggccggcca
gcttggccgctgaagaaaccgagcgccgccgtctaaaaaggtgatgtgtatttgagtaaaacagcttg
cgtcatgcggtcgctgcgtatatgatgcgatgagtaaataaacaaatacgcaaggggaacgcatgaag
gttatcgctgtacttaaccagaaaggcgggtcaggcaagacgaccatcgcaacccatctagcccgcgc
cctgcaactcgccggggccgatgttctgttagtcgattccgatccccagggcagtgcccgcgattggg
cggccgtgcgggaagatcaaccgctaaccgttgtcggcatcgaccgcccgacgattgaccgcgacgtg
aaggccatcggccggcgcgacttcgtagtgatcgacggagcgccccaggcggcggacttggctgtgtc
cgcgatcaaggcagccgacttcgtgctgattccggtgcagccaagcccttacgacatatgggccaccg
ccgacctggtggagctggttaagcagcgcattgaggtcacggatggaaggctacaagcggcctttgtc
gtgtcgcgggcgatcaaaggcacgcgcatcggcggtgaggttgccgaggcgctggccgggtacgagct

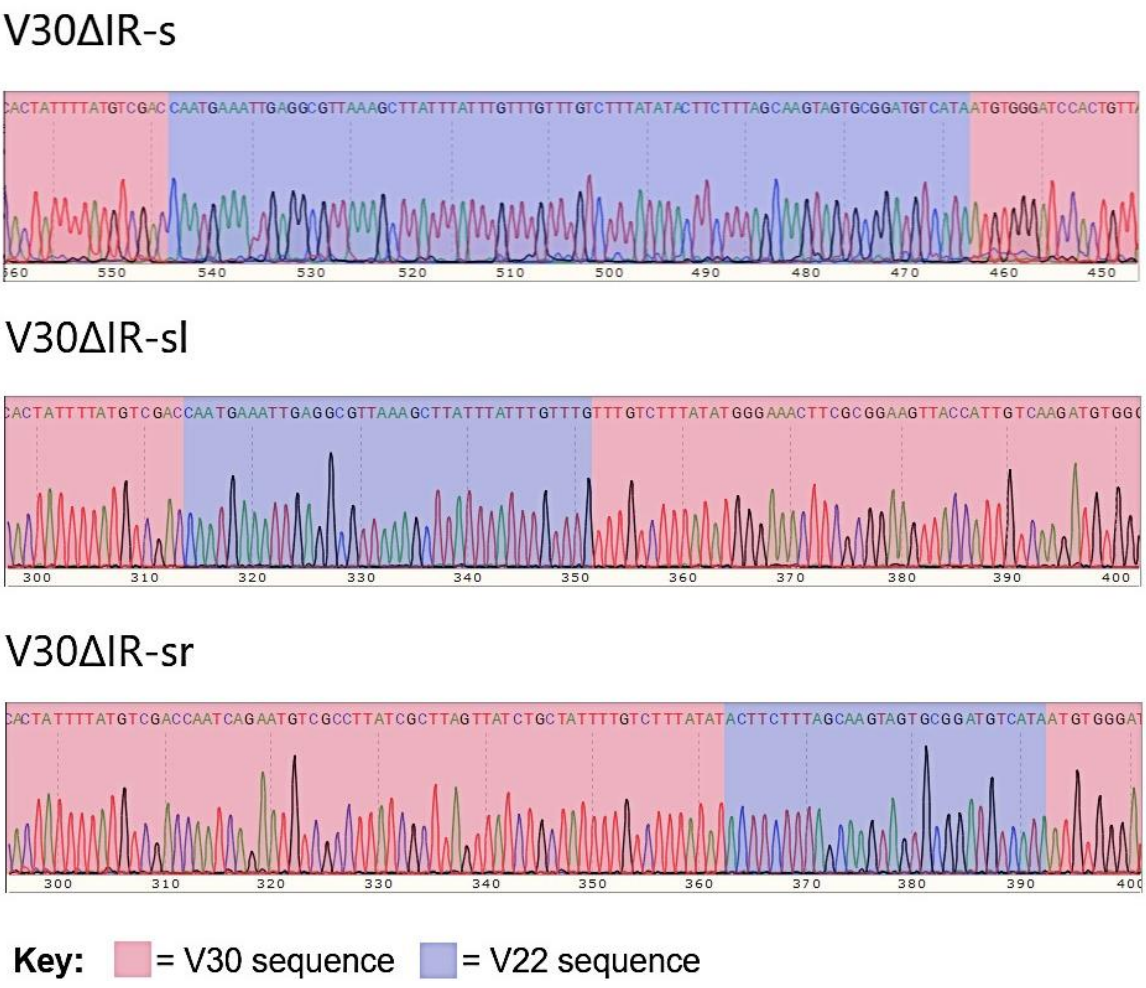
gcccattcttgagtcccgtatcacgcagcgcgtgagctacccaggcactgccgccgccggcacaaccg
ttcttgaatcagaacccgagggcgacgctgcccgcgaggtccaggcgctggccgctgaaattaaatca
aaactcatttgagttaatgaggtaaagagaaaatgagcaaaagcacaaacacgctaagtgccggccgt
ccgagcgcacgcagcagcaaggctgcaacgttggccagcctggcagacacgccagccatgaagcgggt
caactttcagttgccggcggaggatcacaccaagctgaagatgtacgcggtacgccaaggcaagacca
ttaccgagctgctatctgaatacatcgcgcagctaccagagtaaatgagcaaatgaataaatgagtag
atgaatttagcggctaaaggaggcggcatggaaaatcaagaacaaccaggcaccgacgccgtggaat
gccccatgtgtggaggaacgggcggttggccaggcgtaagcggctgggttgtctgccggccctgcaat
ggcactggaacccccaagcccgaggaatcggcgtgacggtcgcaaaccatccggcccggtacaaatcg
gcgcggcgctgggtgatgacctggtggagaagttgaaggccgcgcaggccgcccagcggcaacgcatc
gaggcagaagcacgccccggtgaatcgtggcaagcggccgctgatcgaatccgcaaagaatcccggca
accgccggcagccggtgcgccgtcgattaggaagccgcccaagggcgacgagcaaccagatttttcg
ttccgatgctctatgacgtgggcacccgcgatagtcgcagcatcatggacgtggccgttttccgtctg
tcgaagcgtgaccgacgagctggcgaggtgatccgctacgagcttccagacgggcacgtagaggtttc
cgcagggccggccggcatggccagtgtgtgggattacgacctggtactgatggcggtttcccatctaa
ccgaatccatgaaccgataccgggaagggaagggagacaagcccggccgcgtgttccgtccacacgtt
gcggacgtactcaagttctgccggcgagccgatggcggaaagcagaaagacgacctggtagaaacctg
cattcggttaaacaccacgcacgttgccatgcagcgtacgaagaaggccaagaacggccgcctggtga
cggtatccgagggtgaagccttgattagccgctacaagatcgtaaagagcgaaaccgggcggccggag
tacatcgagatcgagctagctgattggatgtaccgcgagatcacagaaggcaagaacccggacgtgct
gacggttcaccccgattacttttgatcgatcccggcatcggccgttttctctaccgcctggcacgcc
gcgccgcaggcaaggcagaagccagatggttgttcaagacgatctacgaacgcagtggcagcgccgga
gagttcaagaagttctgtttcaccgtgcgcaagctgatcgggtcaaatgacctgccggagtacgattt
gaaggaggaggcggggcaggctggcccgatcctagtcatgcgctaccgcaacctgatcgagggcgaag
catccgccggttcctaatgtacggagcagatgctagggcaaattgccctagcaggggaaaaaggtcga
aaaggtctctttcctgtggatagcacgtacattgggaacccaaagccgtacattgggaaccggaaccc
gtacattgggaacccaaagccgtacattgggaaccggtcacacatgtaagtgactgatataaaagaga
aaaaaggcgatttttccgcctaaaactctttaaaacttattaaaactcttaaaacccgcctggcctgt
gcataactgtctggccagcgcacagccgaagagctgcaaaaagcgcctacccttcggtcgctgcgctc
cctacgccccgccgcttcgcgtcggcctatcgcggccgctggccgctcaaaaatggctggcctacggc
caggcaatctaccagggcgcggacaagccgcgccgtcgccactcgaccgccggcgcccacatcaaggc
accctgcctcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtca
cagcttgtctgtaagcggatgccgggagcagacaagcccgtcagggcgcgtcagcgggtgttggcggg
tgtcggggcgcagccatgacccagtcacgtagcgatagcggagtgtatactggcttaactatgcggca
tcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaa
ataccgcatcaggcgctcttccgcttcctcgctcactgactcgctgcgctcggtcgttcggctgcggc
gagcggtatcagctcactcaaaggcggtaatacggttatccacagaatcaggggataacgcaggaaag
aacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgcgttgctggcgtttttcca
taggctccgccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcgaaacccgacag
gactataaagataccaggcgtttccccctggaagctccctcgtgcgctctcctgttccgaccctgccg
cttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagctcacgctgtag
gtatctcagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaaccccccgttcagcccg
accgctgcgccttatccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactg
gcagcagccactggtaacaggattagcagagcgaggtatgtaggcggtgctacagagttcttgaagtg
gtggcctaactacggctacactagaaggacagtatttggtatctgcgctctgctgaagccagttacct
tcggaaaaagagttggtagctcttgatccggcaaacaaaccaccgctggtagcggtggttttttgtt
tgcaagcagcagattacgcgcagaaaaaaaggatctcaagaagatcctttgatcttttctacggggtc
tgacgctcagtggaacgaaaactcacgttaagggattttggtcatgcattctaggtactaaaacaatt
catccagtaaaatataatattttattttctcccaatcaggcttgatccccagtaagtcaaaaaatagc
tcgacatactgttcttccccgatatcctccctgatcgaccggacgcagaaggcaatgtcataccactt
gtccgccctgccgcttctcccaagatcaataaagccacttactttgccatctttcacaaagatgttgc

tgtctcccaggtcgccgtgggaaaagacaagttcctcttcgggcttttccgtctttaaaaaatcatac
agctcgcgcggatctttaaatggagtgtcttcttcccagttttcgcaatccacatcggccagatcgtt
attcagtaagtaatccaattcggctaagcggctgtctaagctattcgtatagggacaatccgatatgt
cgatggagtgaaagagcctgatgcactccgcatacagctcgataatcttttcagggctttgttcatct
tcatactcttccgagcaaaggacgccatcggcctcactcatgagcagattgctccagccatcatgccg
ttcaaagtgcaggacctttggaacaggcagctttccttccagccatagcatcatgtccttttcccgtt
ccacatcataggtggtccctttataccggctgtccgtcatttttaaatataggttttcattttctccc
accagcttatataccttagcaggagacattccttccgtatcttttacgcagcggtatttttcgatcag
ttttttcaattccggtgatattctcattttagccatttattatttccttcctcttttctacagtattt
aaagataccccaagaagctaattataacaagacgaactccaattcactgttccttgcattctaaaacc
ttaaataccagaaaacagctttttcaaagttgttttcaaagttggcgtataacatagtatcgacggag
ccgattttgaaaccgcggtgatcacaggcagcaacgctctgtcatcgttacaatcaacatgctaccct
ccgcgagatcatccgtgtttcaaacccggcagcttagttgccgttcttccgaatagcatcggtaacat
gagcaaagtctgccgccttacaacggctctcccgctgacgccgtcccggactgatgggctgcctgtat
cgagtggtgattttgtgccgagctgccggtcggggagctgttggctggctggtggcaggatatattgt
ggtgtaaacaaattgacgcttagacaacttaataacacattgcggacgtttttaatgtactgaattaa
cgccgaattaattcgggggatctggattttagtactggattttggttttaggaattagaaatttttatt
gatagaagtattttacaaatacaaatacatactaagggtttcttatatgctcaacacatgagcgaaac
cctataggaaccctaattcccttatctgggaactactcacacattattatggagaaactcgagcttgt
cgatcgactctagctagaggatcgatccgaaccccagagtcccgctcagaagaactcgtcaagaaggc
gatagaaggcgatgcgctgcgaatcgggagcggcgataccgtaaagcacgaggaagcggtcagcccat
tcgccgccaagctcttcagcaatatcacgggtagccaacgctatgtcctgatagcggtccgccacacc
cagccggccacagtcgatgaatccagaaaagcggccattttccaccatgatattcggcaagcaggcat
cgccatgtgtcacgacgagatcctcgccgtcgggcatgcgcgccttgagcctggcgaacagttcggct
ggcgcgagcccctgatgctcttcgtccagatcatcctgatcgacaagaccggcttccatccgagtacg
tgctcgctcgatgcgatgtttcgcttggtggtcgaatgggcaggtagccggatcaagcgtatgcagcc
gccgcattgcatcagccatgatggatactttctcggcaggagcaaggtgagatgacaggagatcctgc
cccggcacttcgcccaatagcagccagtcccttcccgcttcagtgacaacgtcgagcacagctgcgca
aggaacgcccgtcgtggccagccacgatagccgcgctgcctcgtcctggagttcattcagggcaccgg
acaggtcggtcttgacaaaaagaaccgggcgccccctgcgctgacagccggaacacggcggcatcagag
cagccgattgtctgttgtgcccagtcatagccgaatagcctctccacccaagcggccggagaacctgc
gtgcaatccatcttgttcaatccccatggtcgatcgacagatctgcgaaagctcgagagagatagatt
tgtagagagagactggtgatttcagcgtgtcctctccaaatgaaatgaacttccttatatagaggaag
gtcttgcgaaggatagtgggattgtgcgtcatcccttacgtcagtggagatatcacatcaatccactt
gctttgaagacgtggttggaacgtcttcttttttccacgatgctcctcgtgggtgggggtccatctttg
ggaccactgtcggcagaggcatcttgaacgatagcctttcctttatcgcaatgatggcatttgtaggt
gccaccttccttttctactgtccttttgatgaagtgacagatagctgggcaatggaatccgaggaggt
ttcccgatattacccttttgttgaaaagtctcaatagccctttggtcttctgagactgtatctttgata
ttcttggagtagacgagagtgtcgtgctccaccatgttatcacatcaatccacttgctttgaagacgt
ggttggaacgtcttcttttttccacgatgctcctcgtgggtgggggtccatctttgggaccactgtcgg
cagaggcatcttgaacgatagcctttcctttatcgcaatgatggcatttgtaggtgccaccttccttt
tctactgtccttttgatgaagtgacagatagctgggcaatggaatccgaggaggtttcccgatattac
cctttgttgaaaagtctcaatagccctttggtcttctgagactgtatctttgatattcttggagtaga
cgagagtgtcgtgctccaccatgttggcaagctgctctagccaatacgcaaaccgcctctccccgcgc
gttggccgattcattaatgcagctggcacgacaggtttcccgactggaaagcgggcagtgagcgcaac
gcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtat
gttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatgacatgattacgaattcga
gctccaaagcggccatccgattaatattaccggatggccgtgcccccaaaaaaaactagtccccacgc
actattttatgtcgaccaatgaaattgaggcgttaaagcttatttatttgtttgtttgtctttatata
cttctttagcaagtagtgcggatgtcataatgtgggatccactgttaaacgagttcccagactctgtt
catgggtttcgttgtatgcttgctataaaatacttgcaggttattgagtccacttatgagcccaatac

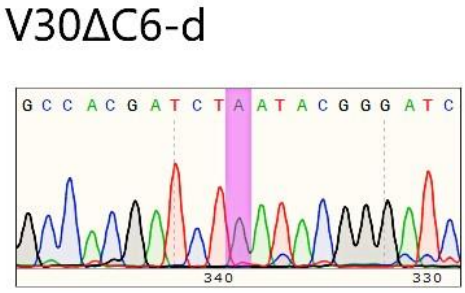
tttgggccacgatttaattcgagatcttatttctgttattagagcccgggattatgtcgaagcgaccc
gccgatataatcatttccacgcccgcctcgaaggtgcggagaaggctgaacttcgacagcccgttcac
cagccgtgctgttgtccccattgtcccaggcacaaacaagcgacgatcatggacgtacaggcccatgt
atcgaaagcccaggatgtacagaatgtacagaagccctgatgtccctcgaggttgtgaaggcccatgt
aagatccagtcgtatgaacagcgcgatgatgtgaagcacaccggtatagttaggtgtatcagtgacat
tacaaagggtcctggtctgacccatcgcacgggtaaaaggttcgtggtgaagtctgtctatatactcg
gtaaggtatggatggatgagaacatcaagaagcagaaccacactaacaatgttatgtttttcttagtc
cgagatagaaggccctatggaacaagcccaatggactttgggcaggtgtttaacatgtttgataacga
gcccagtacagccacggtgaagaacgatcttcgtgatcgattccaggtgttgcgcaagtttacagcca
ccgttgttggtggtccttctggtcttaaggagcaggctcttgtgaagcgtttcttcaggttaaatagc
catgtaacctataaccatcaggaagcagccaagtacgagaaccataccgagaatgcgttgttgttgta
tatggcatgtacccatgcctctaatcctgtgtatgccacgttaaaaatacgtgtatatttttacgact
cagtatcgaattaataaatattaaatttttatttcatgattctccttaacttggatagtgtttacaatt
acatcgtaaagtacatgatccactgctcgaattacattgttaattgaaattacacctaatacatctaa
aaacttaagaacttgggtcctaaatacgcttaagaaacgaccagtctgaggctgtaaggtcgtccaga
cttggaagttgagaaaacatttgtgaatccccaattccttccggaggttgtggttgaaccgtatctgg
atcgatatgatgtcgtggttcgtgtggaatggccgtttctcgtggttgataatcttgaaatatagggg
atttgttatctcccaaatatagacgccattctgggcctgatgagcagtgatgagttccccggtgcgta
aatccatggttgctgcagttgatatggagataatatgtacacccgcaatcaagatcaactcgcctacg
cctgactactctcttcttggctattcgatgctggaccttgattgggacttgagtacagtggctggttg
agggtgatgaagatcgcattctttatagtccaagccttcagggatgcattcttatcctcatcgagata
ttctttatatgaggatgttgggcctggattgcagaggaagatagtgggaatgccgcctttaatttgaa
ttggcttcccgtacttcgtgttgctttgccagtccctttgggccccaaagaattccttccagtgcttt
aggtaatgcggatctacgtcatcaatgacgttataccaagcgtcattagaatatacctttggacttaa
gtctaaatgaccacataagtaattatgtggccctaatgatctagcccaaatagttttgccggaacgac
tatcgccctctataacaatactattaggtctccaaggccgcgcagcggcaccacacacattctcagag
acccactcttcaagttcctcagggacttgatcaaaagaagaagaagaaaagggagaaatataaggctt
tgggggagcctggaaaatcctatctaaattggtctttaaattatgatactgaaaaataaaatctttag
ggagtttttcccttataatagccatagcggcttcagcggaatctgcgtttaaagcgtcggcgcatgcg
tcgttagcatttttgcagcctcctctagcacttctgccgtcgacctggaattctccccattcaagtgt
gtctccatccttgtcgatgtaggacttgacgtcggagctcgatttagctccctgaatgtttggatgga
aatgtgctgacctgtttggggataccaggtcgaagaatcggtgattcttgcagttgaattttccctcg
aactgaataagcacgtggagatgaggttccccattttcgtgtagctctctgcagattttgatgtattt
tttattggtgggtgtttcaaggtttaataattgagagagtgactcttcctttgtgagagagcattttg
gatatgtgacgaaatagtttttggaatttatttgaaagcgcttcggaggagccatttggtcaatgggt
accgattgactctcaaattgcattctcctggtatatcggtacccaatatatagtgggtaccgaatggc
acaattgtaatttgtggcaaagtaattggaatttcaaaattcaaactcctaaagcggccatccgatta
atattaccggatggccgtgcccccaaaaaaaactagtaaaaaagcttggatcctctagagtcgacctg
caggcatgcaagcttggcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttaccc
aacttaatcgccttgcagcacatccccctttcgccagctggcgtaatagcgaagaggcccgcaccgat
cgcccttcccaacagttgcgcagcctgaatggcgaatgctagagcagcttgagcttggatcagattgt
cgtttcccgccttcagtttaaactatcagtgtttgacaggatatattggcgggtaaacctaagagaaa
agagcgtttattagaataacggatatttaaaagggcgtgaaaaggtttatccgttcgtccatttgtat
gtg

Appendix (A) 2

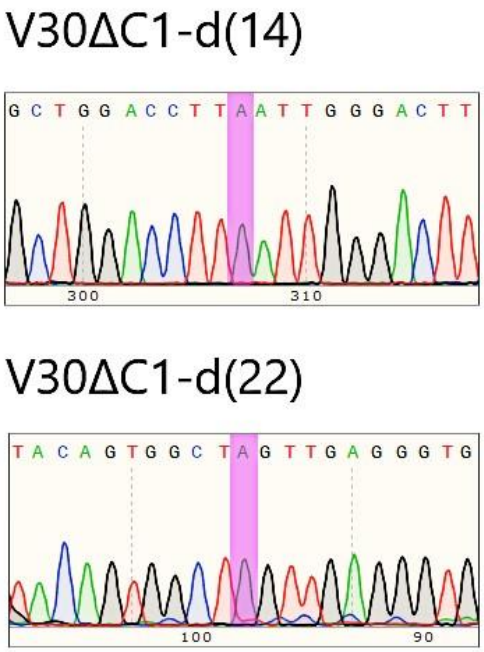
Sequencing chromatograms



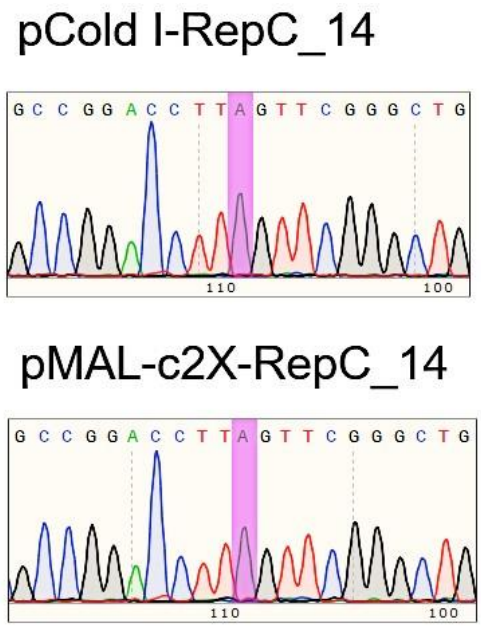
A2 Fig. 1 Sanger sequence chromatograms generated for V30ΔIR-s, V30ΔIR-sl and V30ΔIR-sr IR swap mutants used in this study. Partial 3' IR sequence shown with sequence identity highlighted according to colour key. Expected sequence identities were confirmed.



A2 Fig. 2 Sanger sequence chromatogram generated for V30ΔC6-d C6 deletion mutant used in this study. Partial sequence shown with mutated sequence highlighted in magenta. Expected sequence mutation was confirmed.

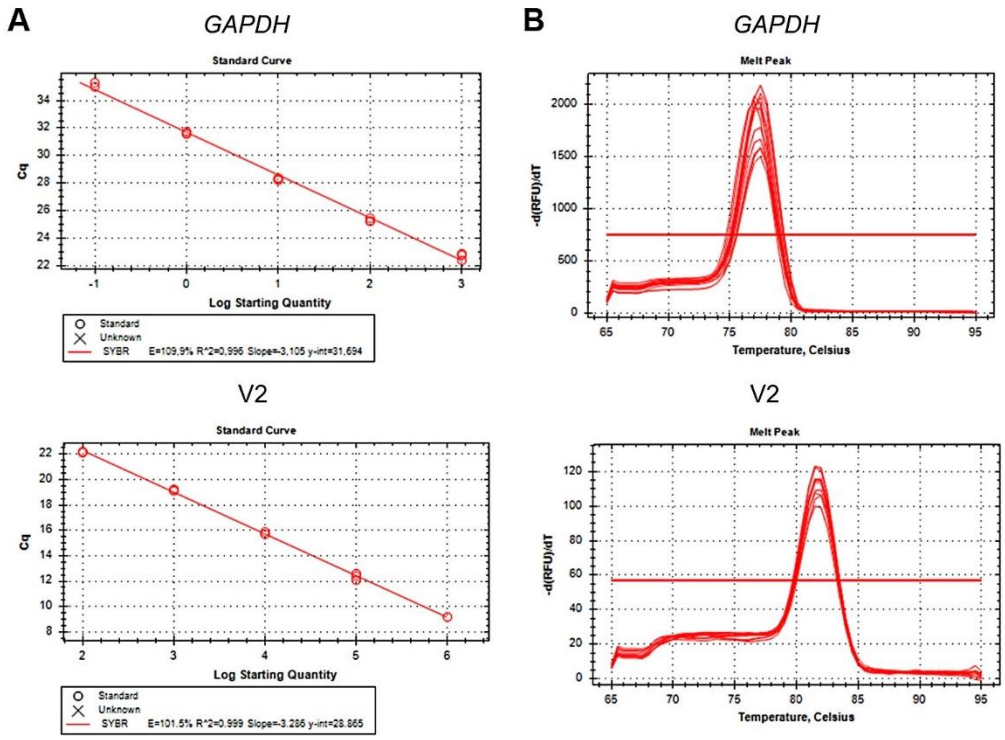


A2 Fig. 3 Sanger sequence chromatogram generated for V30ΔC1-d(14) and V30ΔC1-d(22) C1 deletion mutants used in this study. Partial sequence shown with mutated sequences highlighted in magenta. Expected sequence mutations were confirmed.

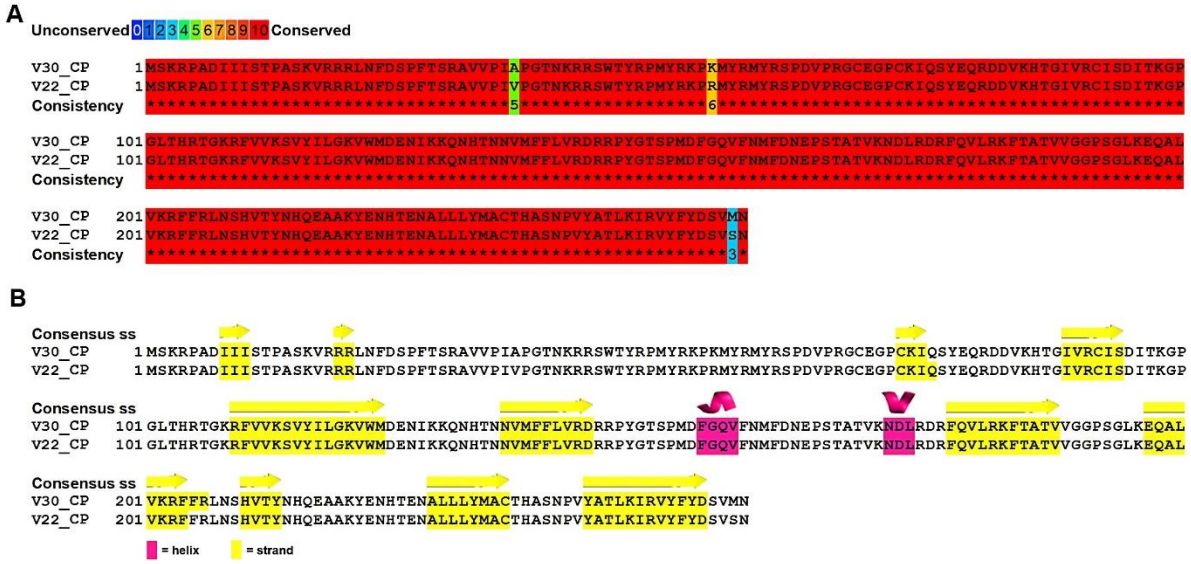


A2 Fig. 4 Sanger sequence chromatogram generated for pCold I-RepC_14 and pMAL-c2X-RepC_14 expression vector deletion mutants used in this study. Partial sequence shown with mutated sequences highlighted in magenta. Expected sequence mutations were confirmed.

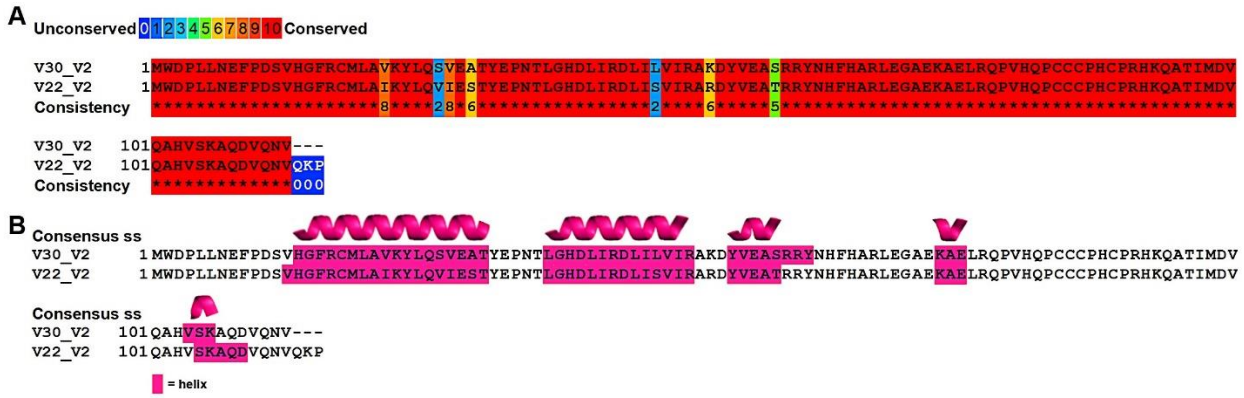
Supplementary 2 (S2) Figures



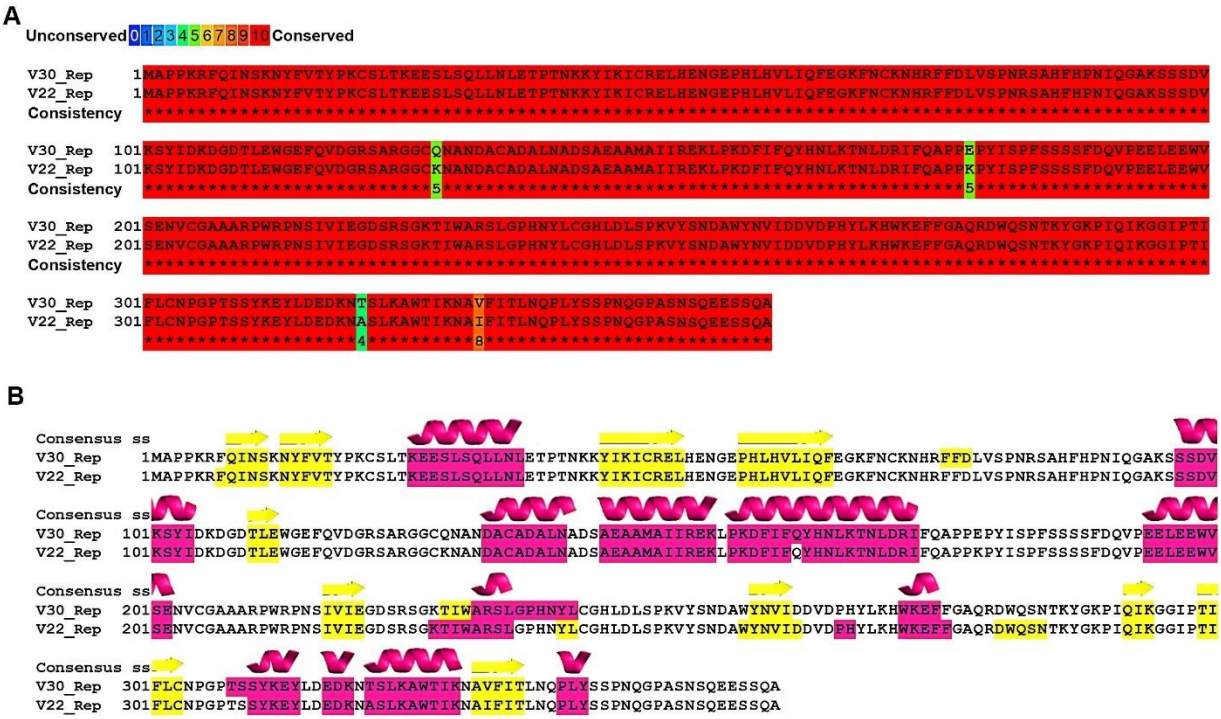
S2 Fig. 2.1 qPCR primer testing. **A**, standard curves generated for internal control glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and viral V2 ORF (V2) primers using a 10-fold dilution series of 100ng total DNA extracted from *Nicotiana benthamiana* inoculated with ToCSV V30. **B**, corresponding qpCR melting curves showing single peaks for both products.



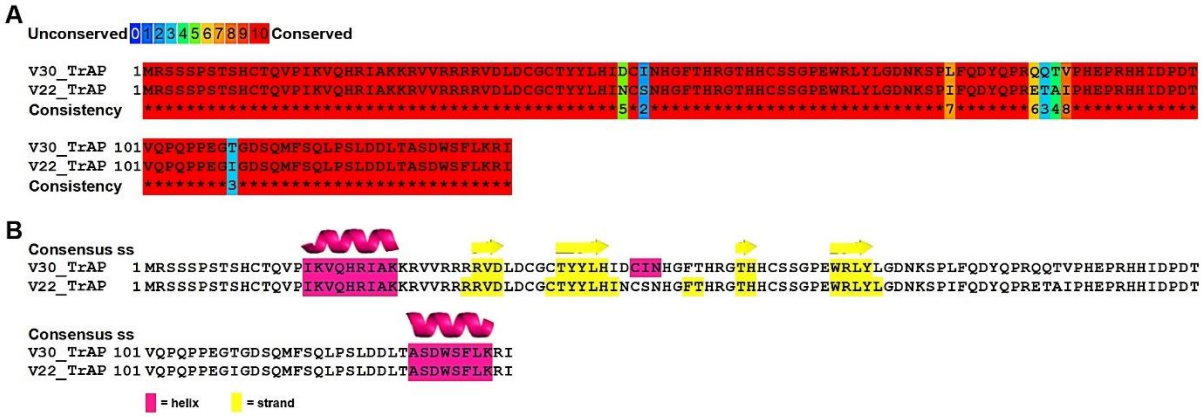
S2 Fig. 2.2 ToCSV V30 and V22 coat protein (CP) aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



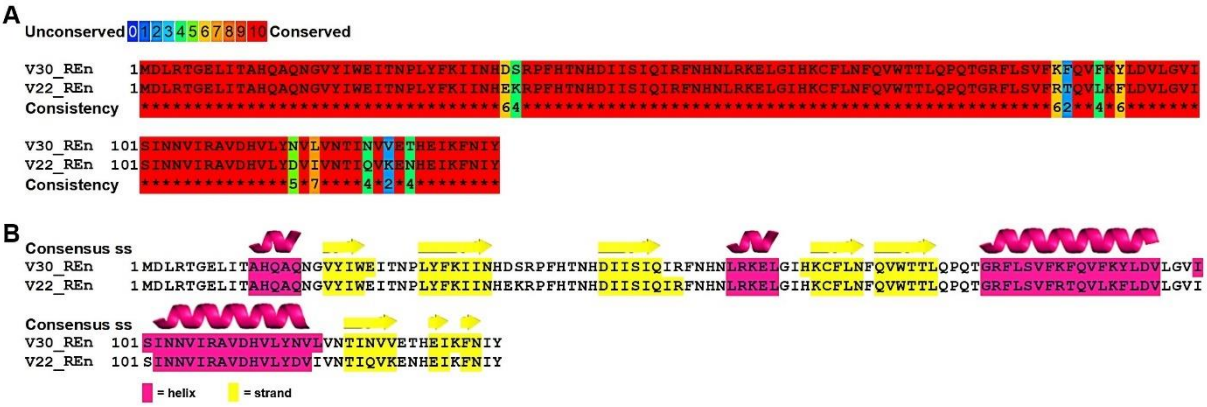
S2 Fig. 2.3 ToCSV V30 and V22 V2 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



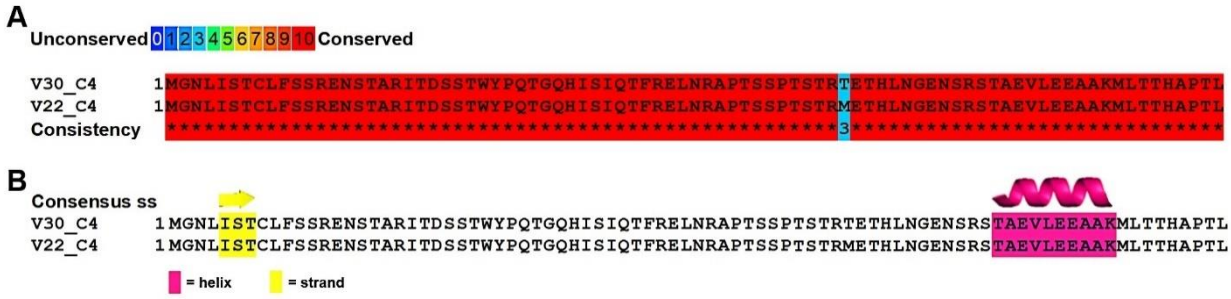
S2 Fig. 2.4 ToCSV V30 and V22 replication-associated protein (Rep) aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



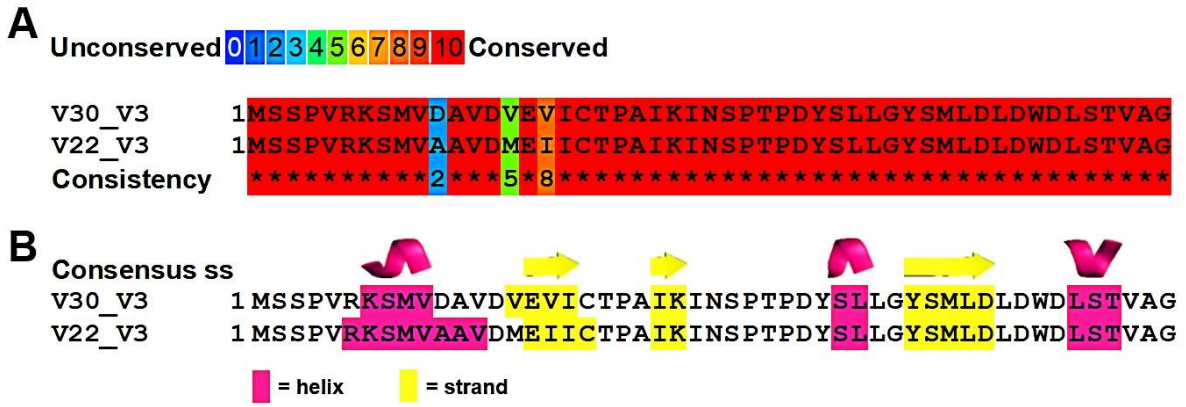
S2 Fig. 2.5 ToCSV V30 and V22 transcriptional activator protein (TrAP) aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



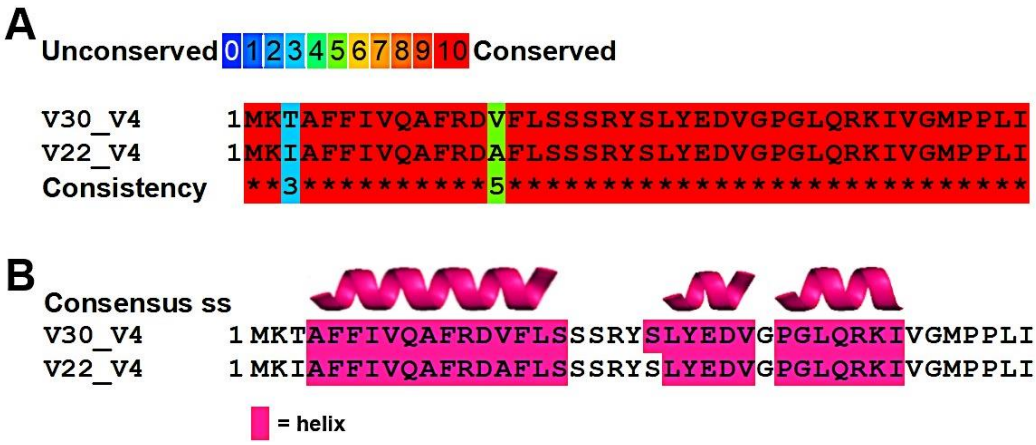
S2 Fig. 2.6 ToCSV V30 and V22 replication enhancer protein (REn) aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



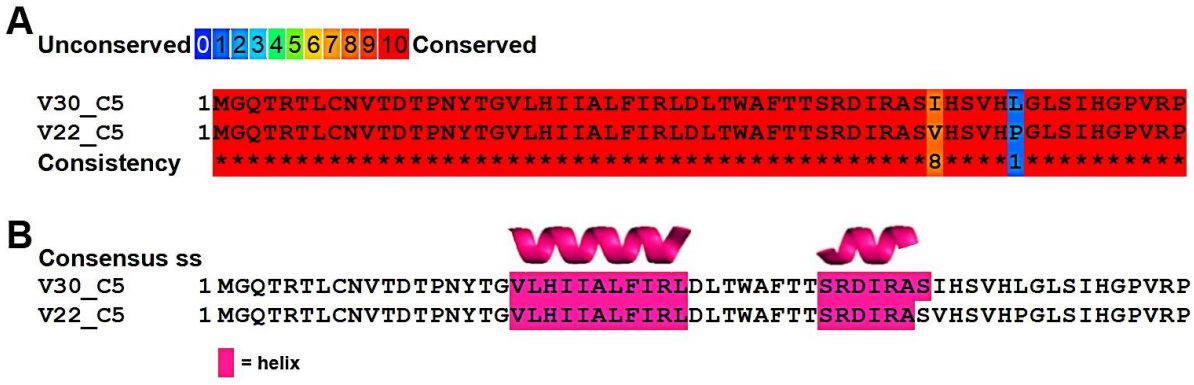
S2 Fig. 2.7 ToCSV V30 and V22 C4 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



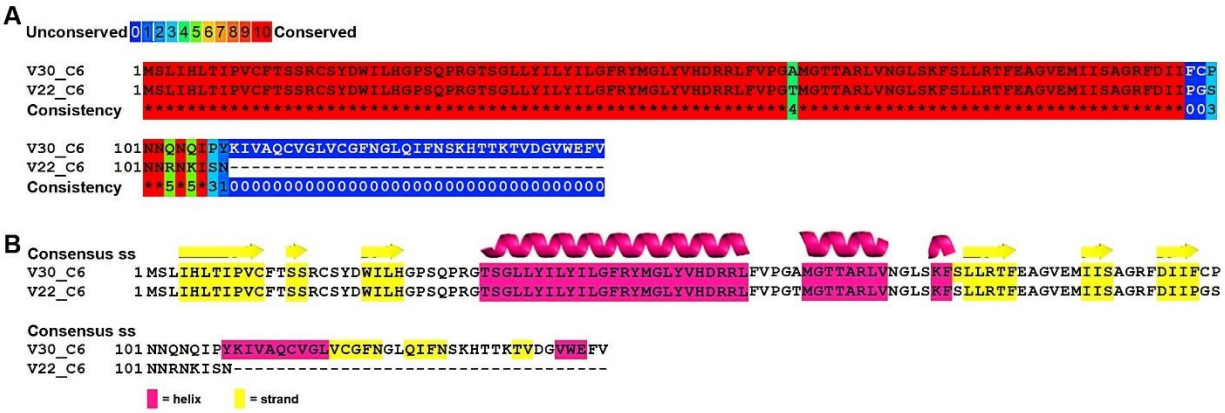
S2 Fig. 2.8 ToCSV V30 and V22 V3 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



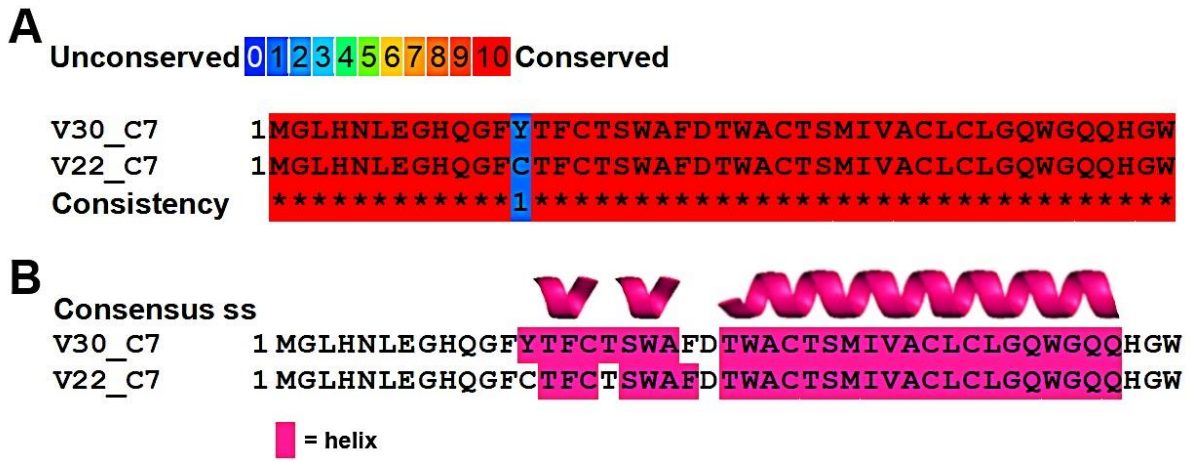
S2 Fig. 2.9 ToCSV V30 and V22 V4 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



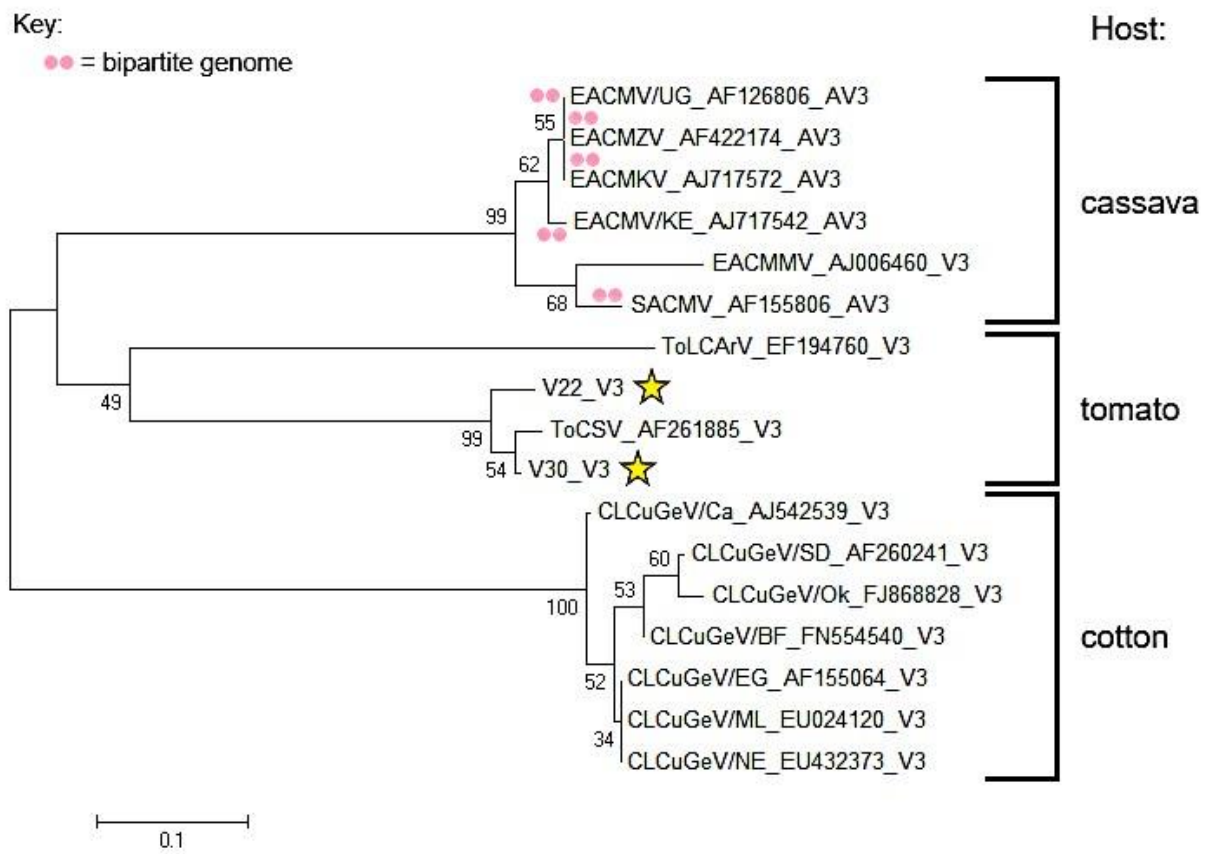
S2 Fig. 2.10 ToCSV V30 and V22 C5 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



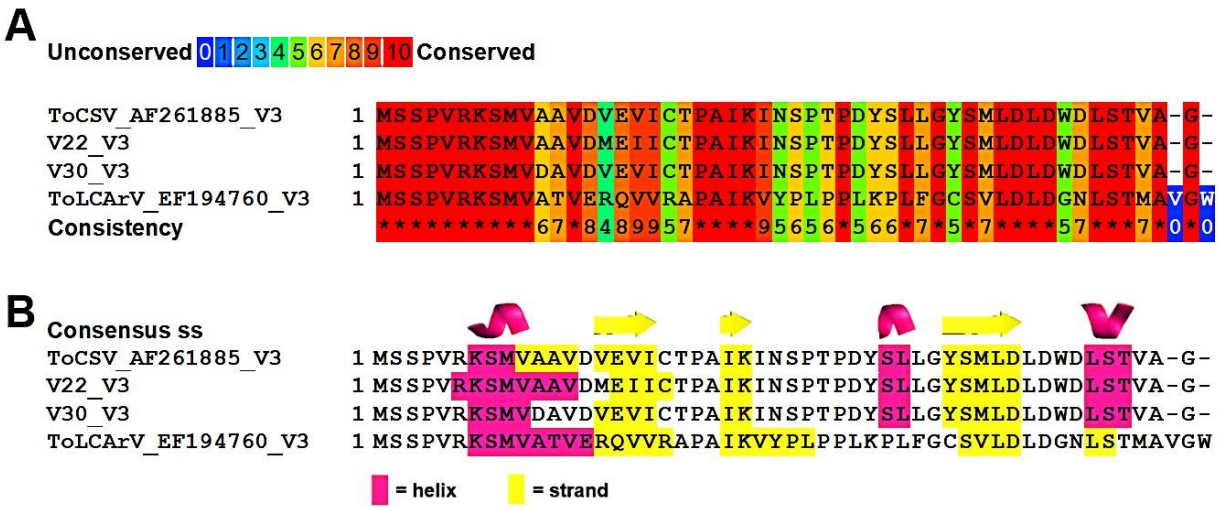
S2 Fig. 2.11 ToCSV V30 and V22 C6 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



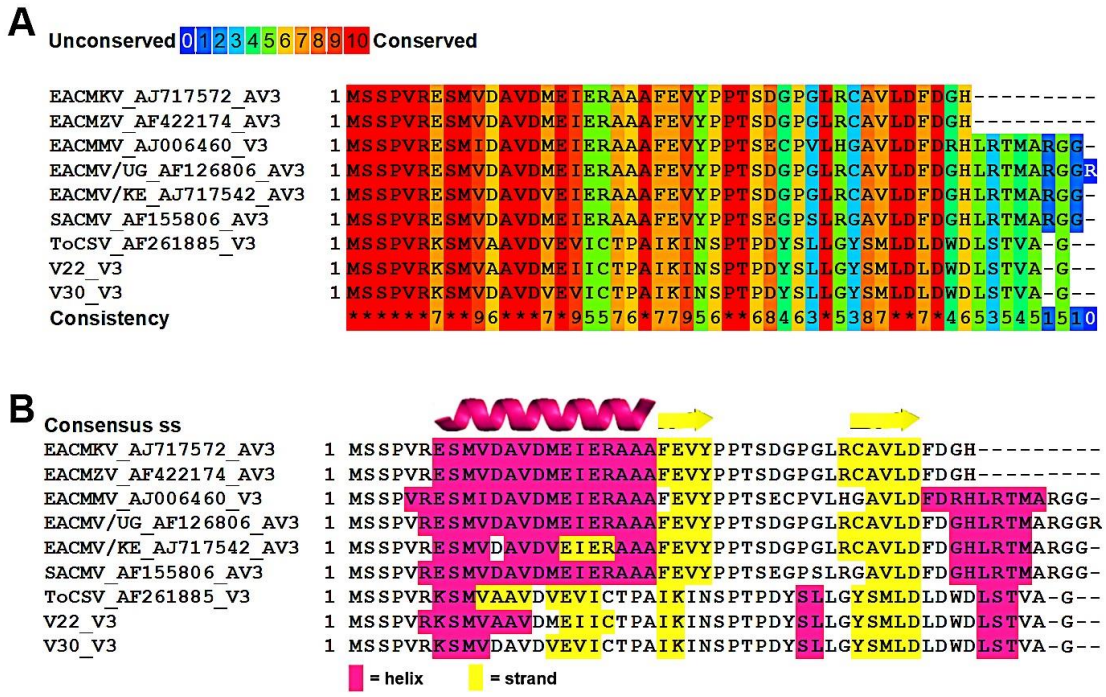
S2 Fig. 2.12 ToCSV V30 and V22 C7 aa sequence pairwise alignment. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.



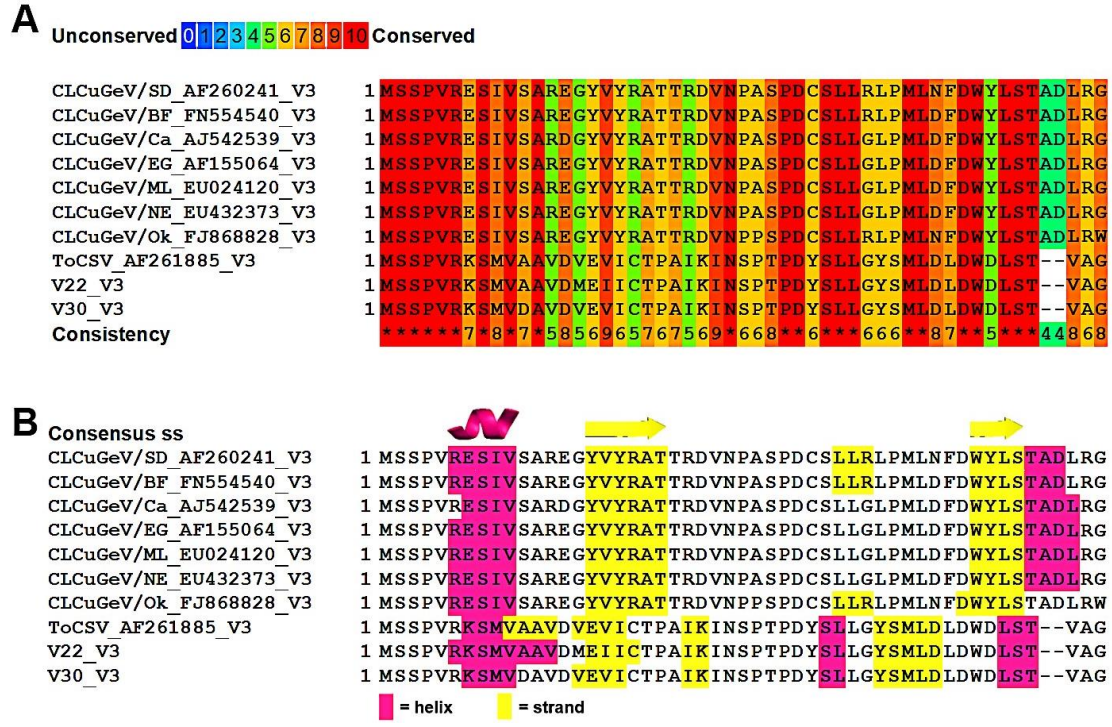
S2 Fig. 2.13 Phylogenetic tree constructed using full length V3/AV3 aa sequences of ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. Alignment done using the neighbour-joining method (1,000 replications). V30 and V22 indicated using stars. Details of all isolate sequences used in S2 Table 2.3.



S2 Fig. 2.14 V3 multiple aa sequence alignment for ToCSV V30, V22, exemplar isolate (AF261885), and ToLCArV (EF914760). **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons.

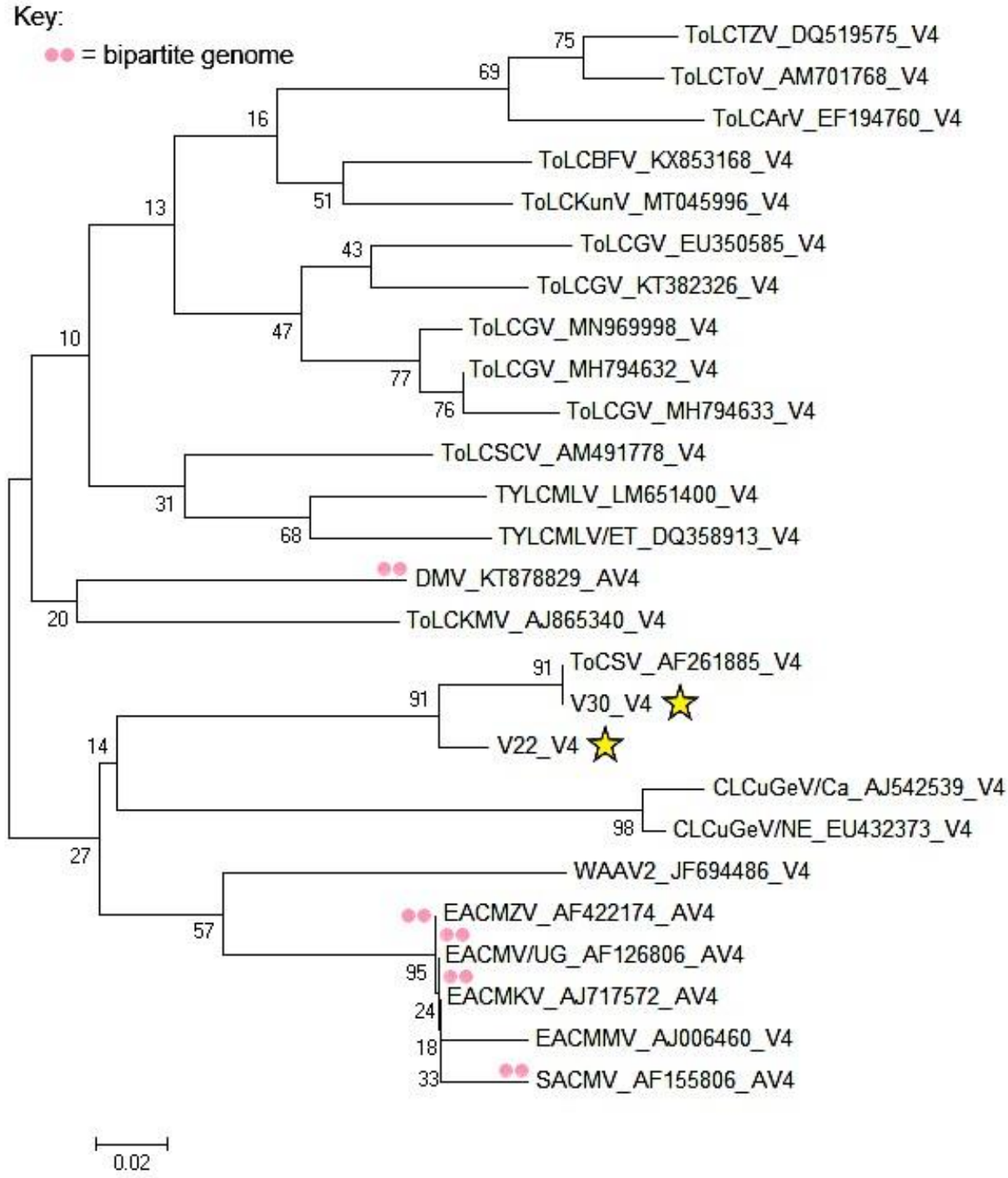


S2 Fig. 2.15 V3/AV3 multiple aa sequence alignment for ToCSV V30, V22, exemplar isolate (AF261885), and selected African cassava-infecting begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.

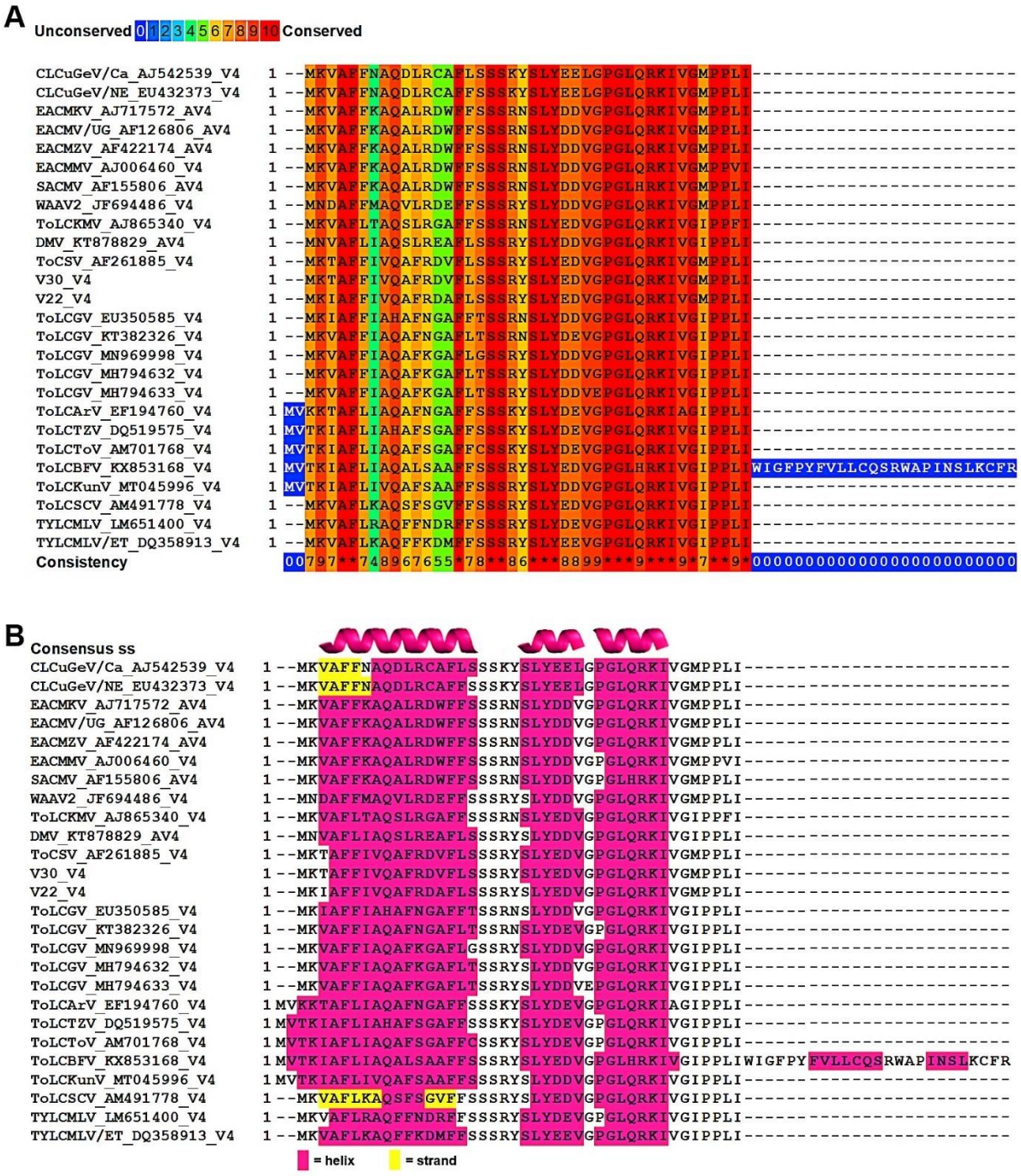


S2 Fig. 2.16 V3 multiple aa sequence alignment for ToCSV V30, V22, exemplar isolate (AF261885), and selected African cotton-infecting begomoviruses sharing ≥75 % DNA-A sequence identity with

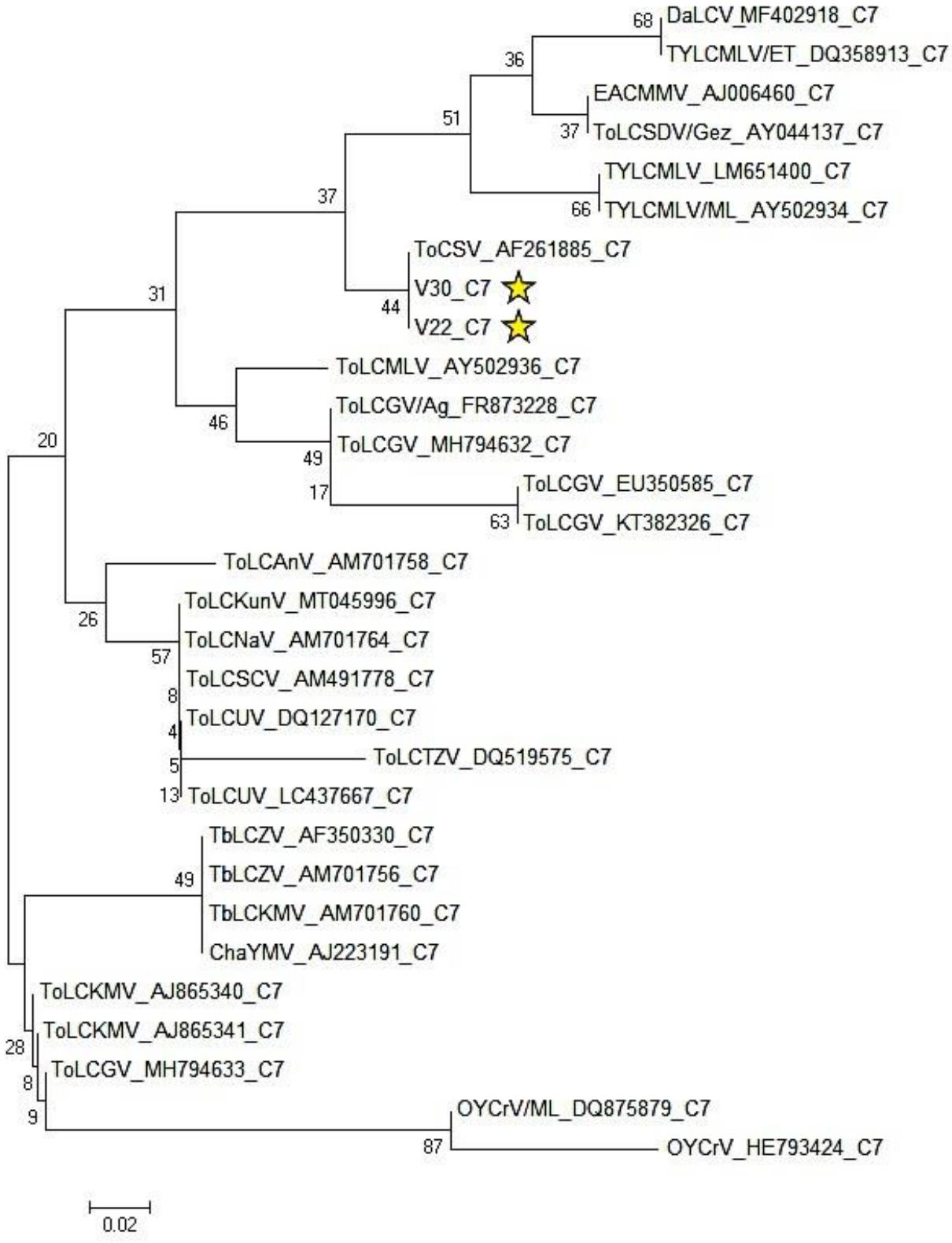
ToCSV. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.



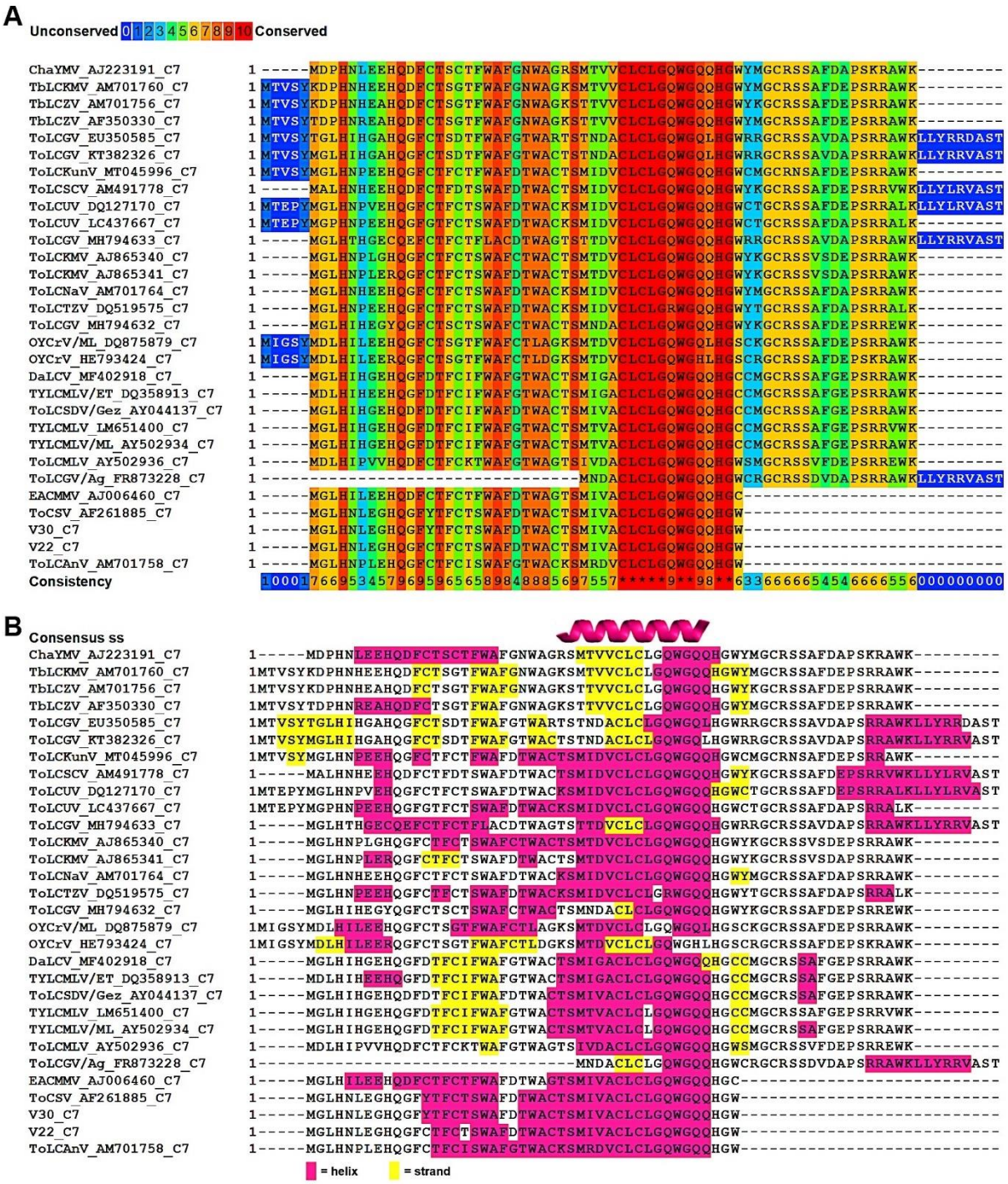
S2 Fig. 2.17 Phylogenetic tree constructed using full length V4/AV4 aa sequences of ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. Alignment done using the neighbour-joining method (1,000 replications). V30 and V22 indicated using stars. Details of all isolate sequences used in S2 Table 2.3.



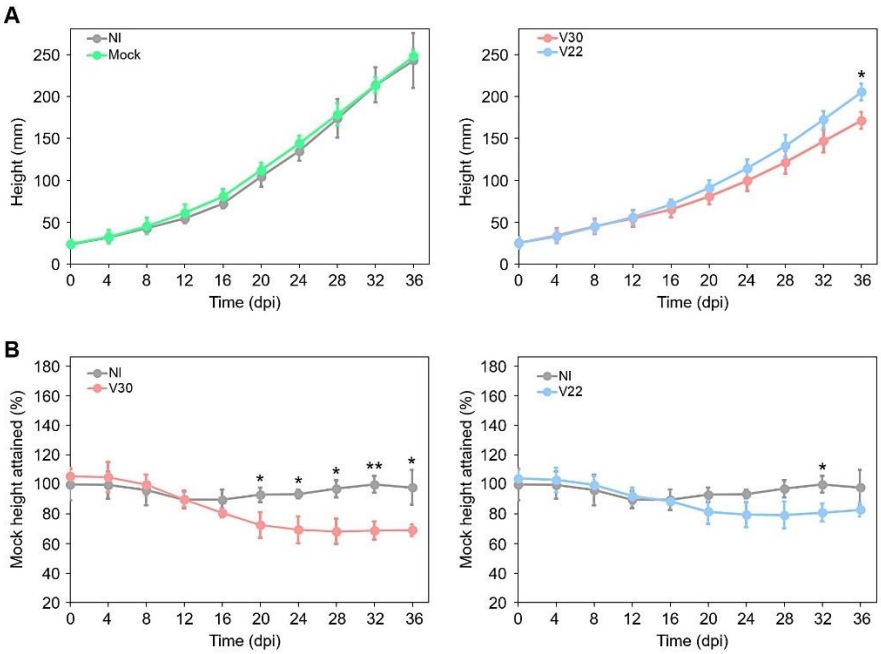
S2 Fig. 2.18 V4/AV4 multiple aa sequence alignment for ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.



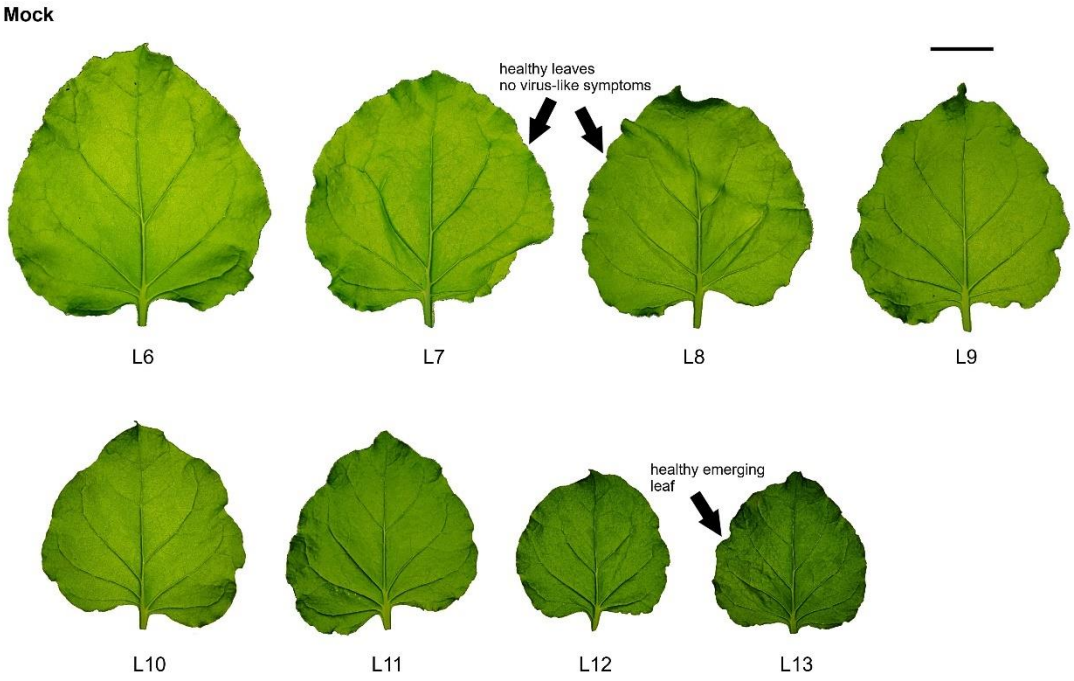
S2 Fig. 2.19 Phylogenetic tree constructed using full length C7 aa sequences of ToCSV V30, V22, and selected African monopartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. Alignment done using the neighbour-joining method (1,000 replications). V30 and V22 indicated using stars. Details of all isolate sequences used in S2 Table 2.3.



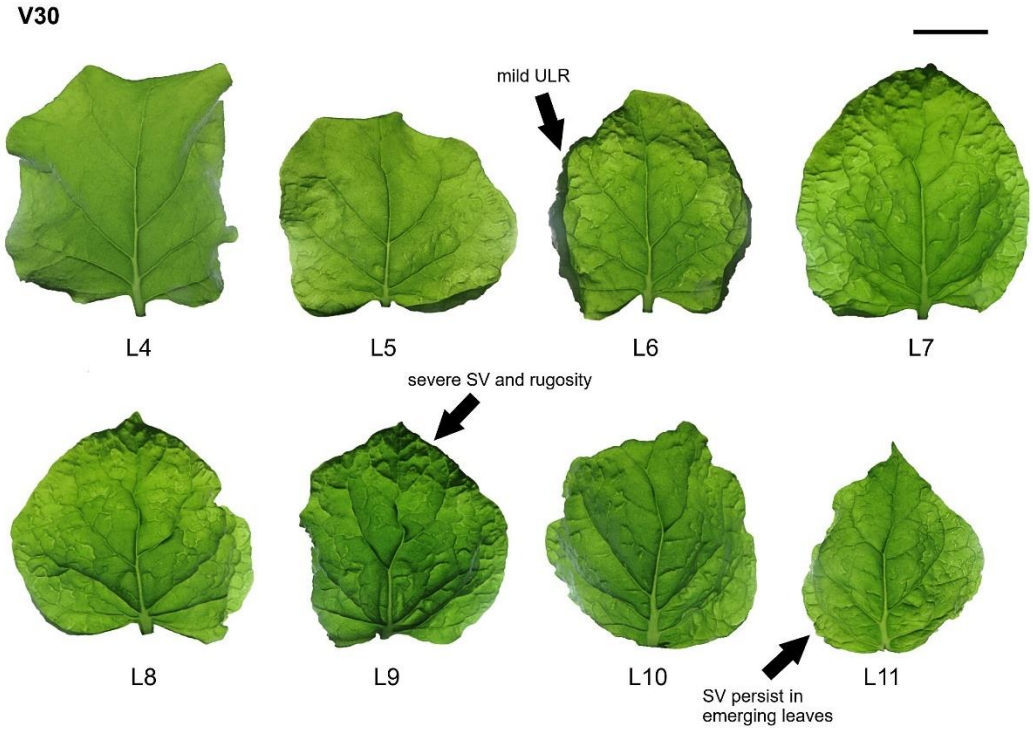
S2 Fig. 2.20 C7 multiple aa sequence alignment for ToCSV V30, V22, and selected African monopartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.



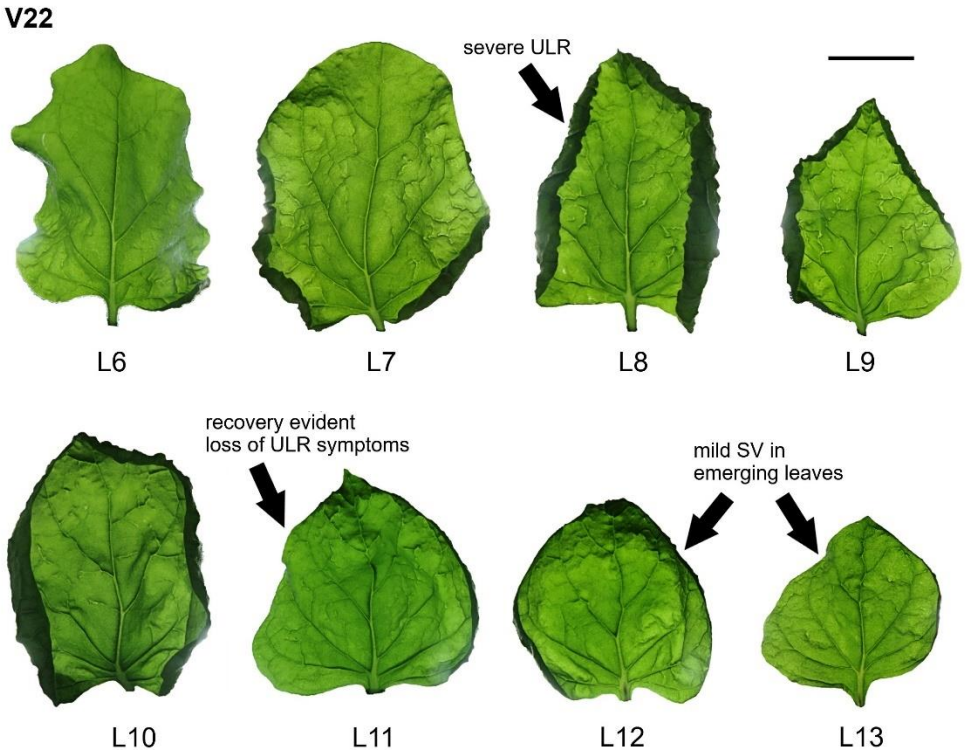
S2 Fig. 2.21 Infection of *Nicotiana benthamiana* with ToCSV wild type infectious clones in single infection. **A**, Heights in non-inoculated plants (NI) and plants inoculated with mock (*Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300), V30, and V22 at four-day intervals from 0 to 36 dpi. **B**, mock height attained in NI, V30 and V22 at four-day intervals from 0-36 dpi. Mock height attained calculated using: (treatment height/mock height)*100. Bars represent mean ± SD. Student's t-test levels of significance: *, $P < 0.05$; **, $P < 0.01$.



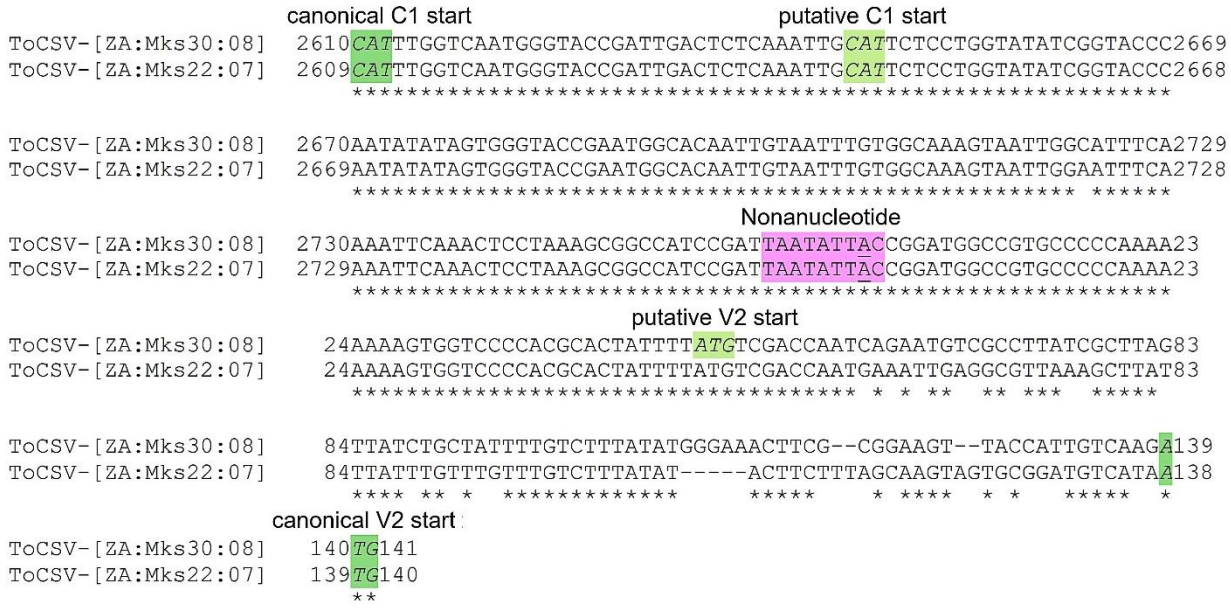
S2 Fig. 2.22 Representative *Nicotiana benthamiana* plant inoculated with mock (*Agrobacterium tumefaciens* C58C1 carrying empty pCAMBIA2300), dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



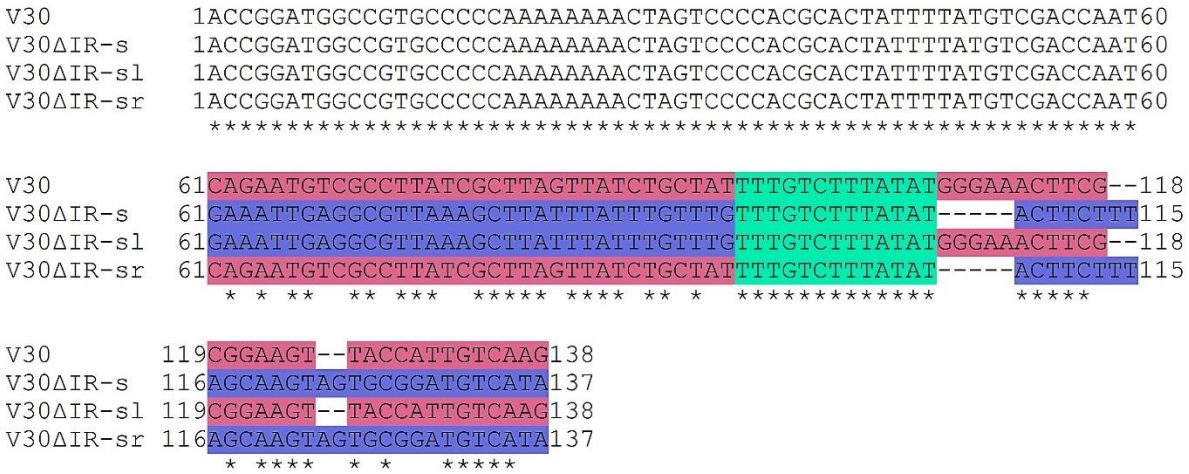
S2 Fig. 2.23 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



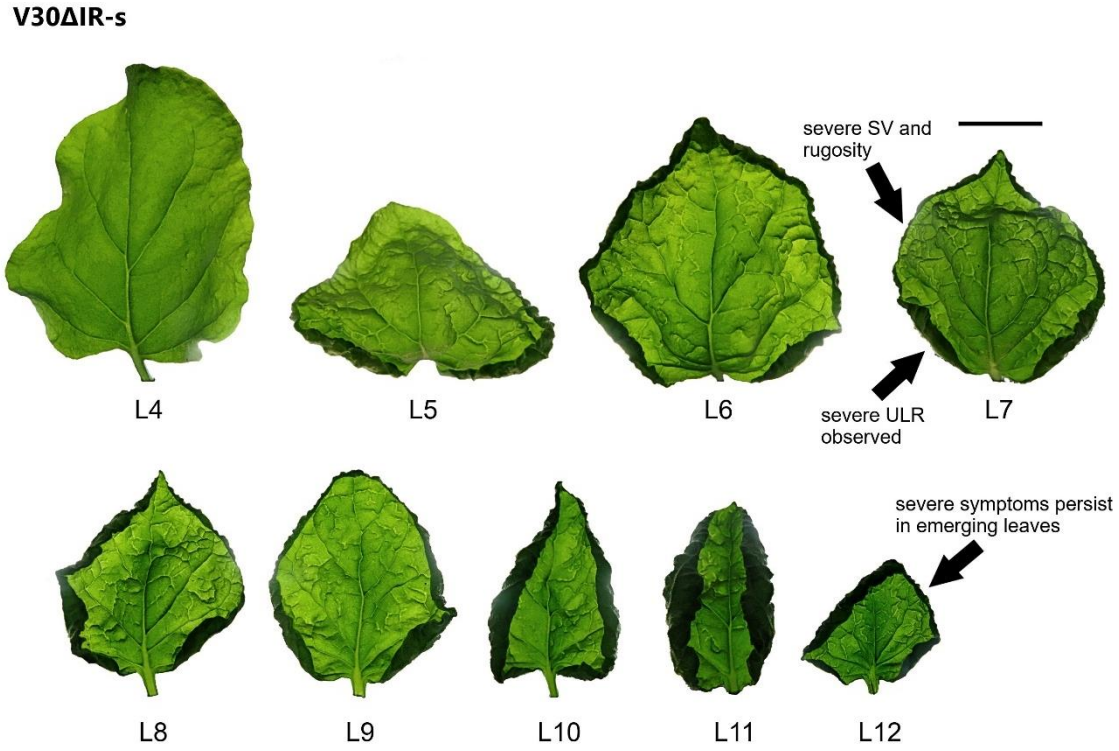
S2 Fig. 2.24 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V22, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



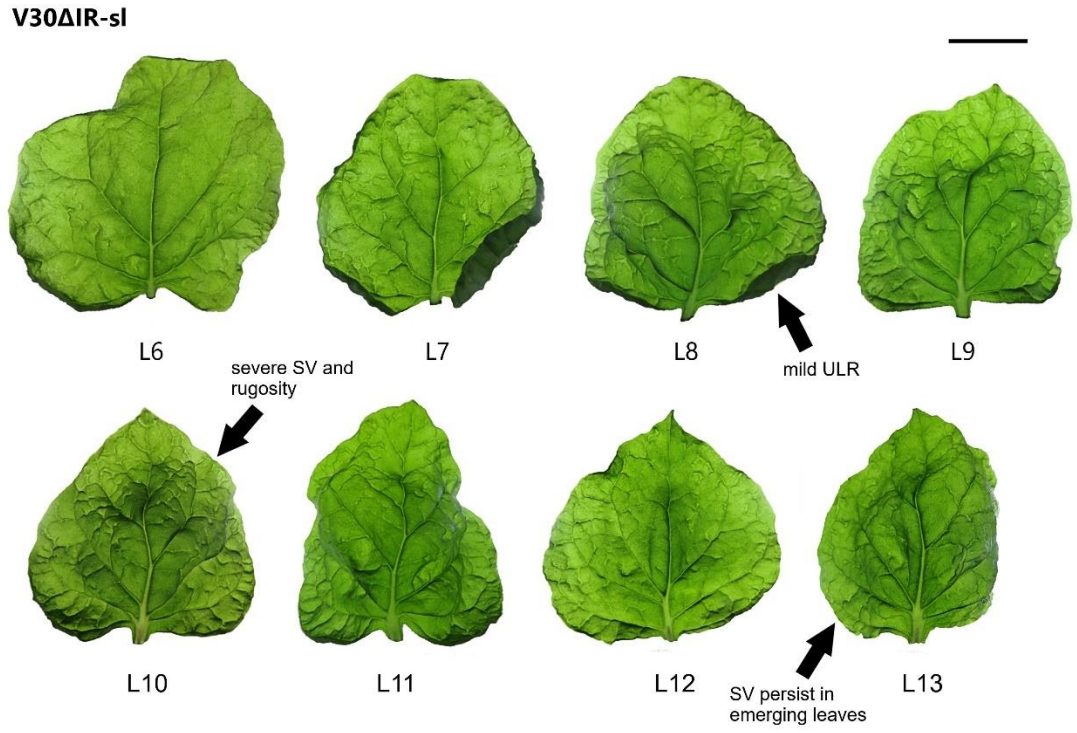
S2 Fig. 2.25 Pairwise full-length intergenic region (IR) nt sequence alignment for ToCSV-[ZA:Mks30:08] and ToCSV-[ZA:Mks22:07]. Alignment done using MUSCLE alignment program with manual adjustments. Nonanucleotide highlighted in magenta with genome nt position 1 underlined. Italicised V2 and C1 canonical start codons highlighted in dark green, putative start codons highlighted in light green. Hyphens indicate alignment-generated gaps. Asterisks indicate sequence conservation.



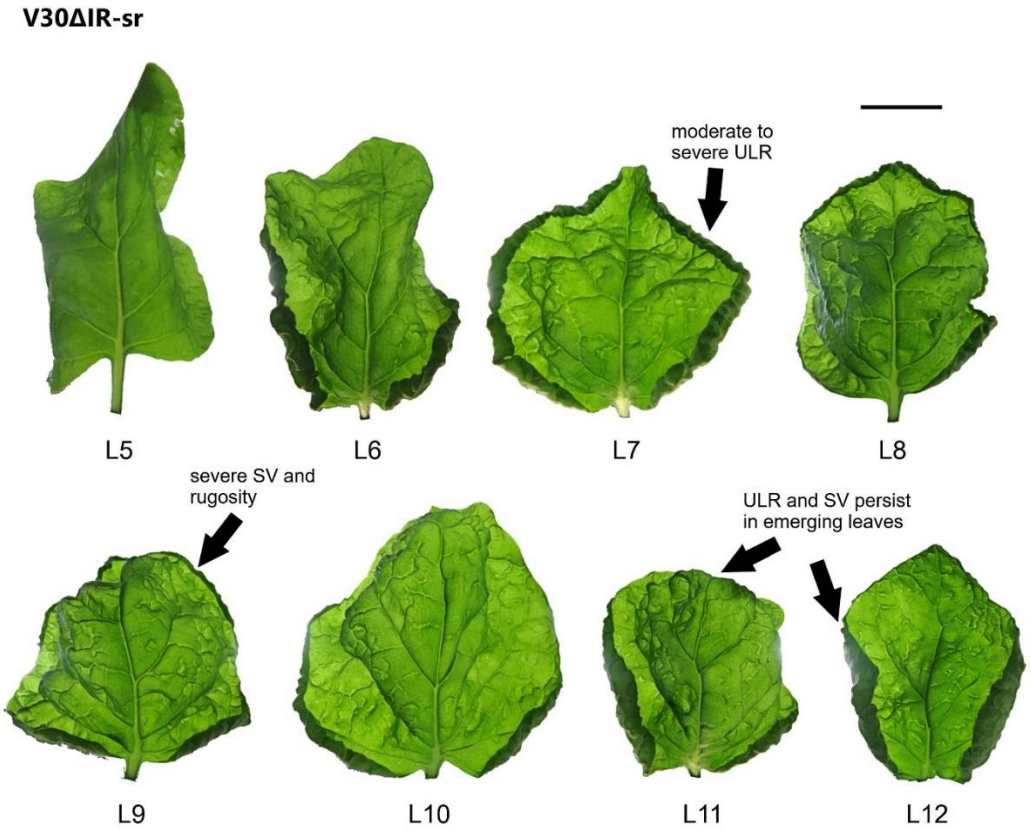
S2 Fig. 2.26 Partial ToCSV intergenic region (IR) nt sequence alignment. V30 and three V30 IR-swap mutants: V30ΔIR-s, V30ΔIR-sl, and V30ΔIR-sr. Alignment done using MUSCLE alignment program with manual adjustments. Swap sequence fragments indicated: V30 sequence in red, V22 sequence in blue. Conserved region containing TATA box in green. Hyphens indicate alignment-generated gaps. Asterisks indicate sequence conservation.



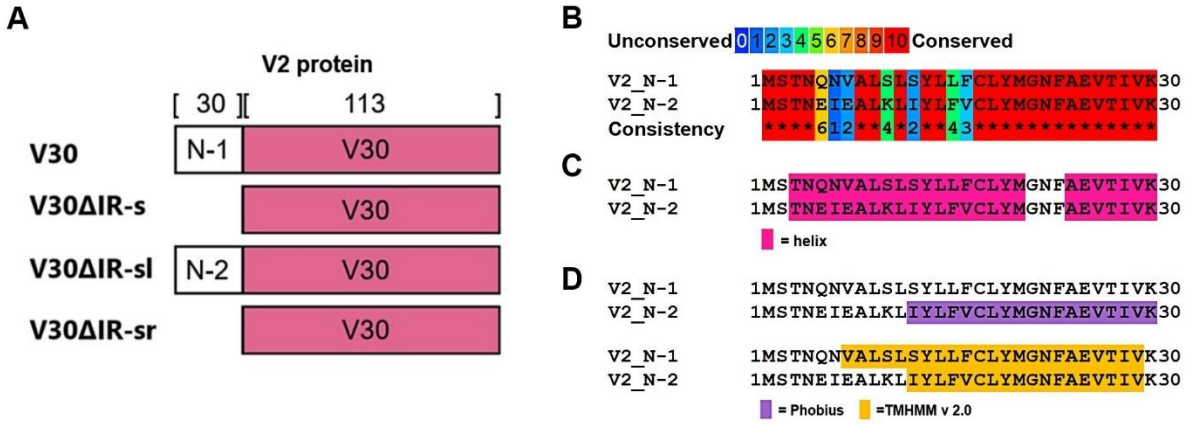
S2 Fig. 2.27 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30ΔIR-s, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



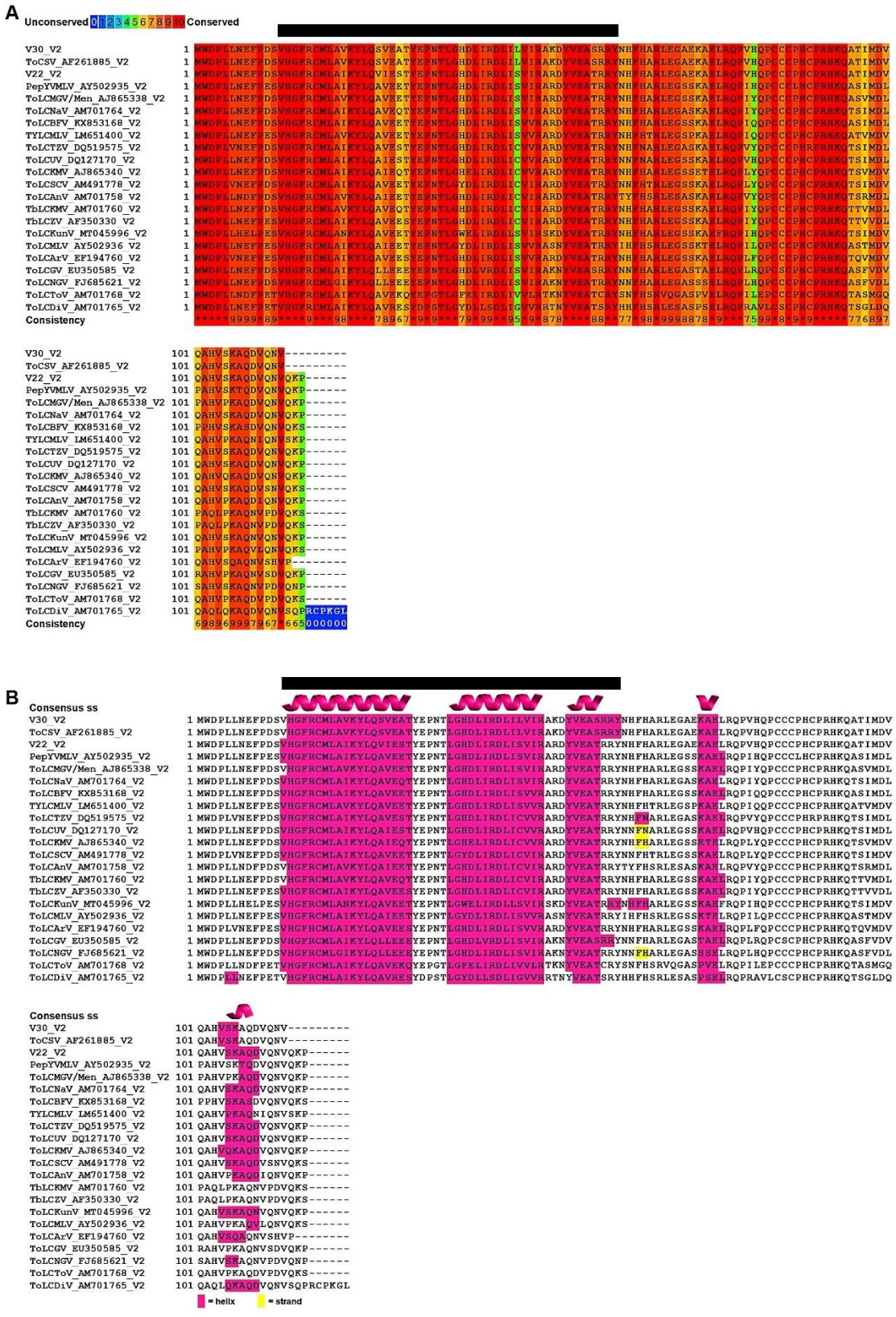
S2 Fig. 2.28 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30ΔIR-sl, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



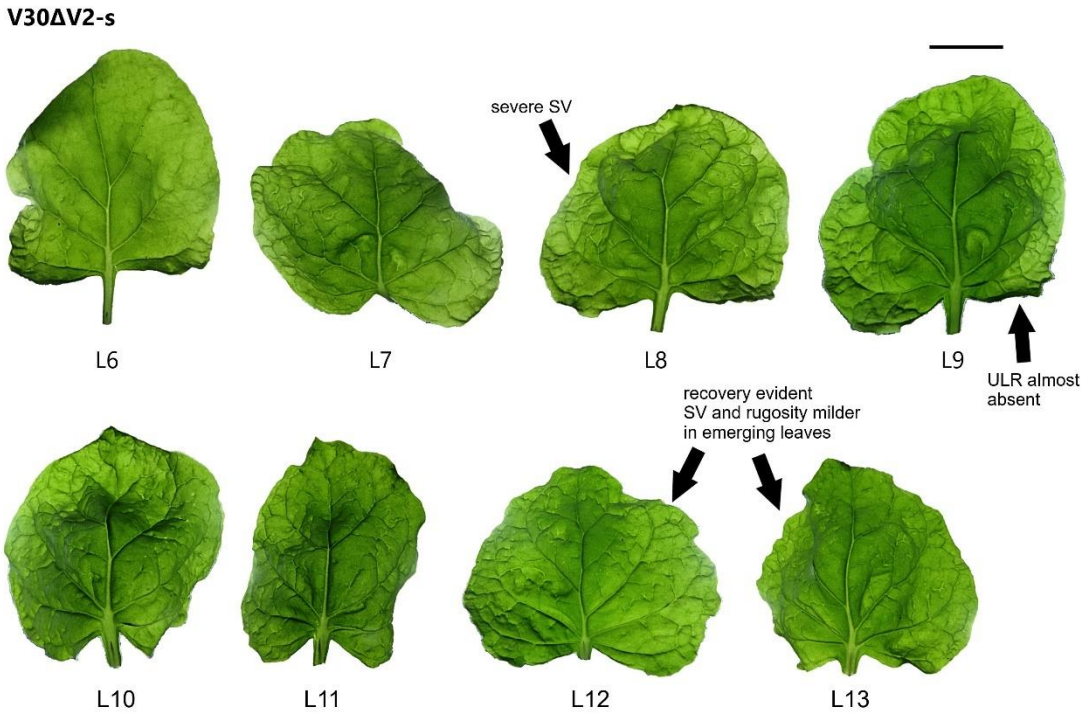
S2 Fig. 2.29 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30ΔIR-sr, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



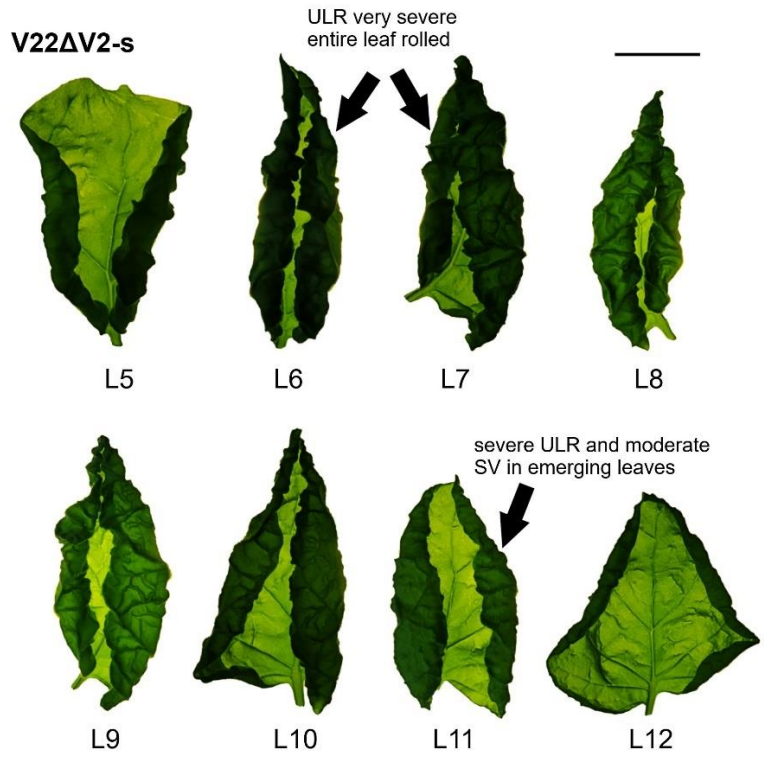
S2 Fig. 2.30 Sequence analysis of V2 protein putative N-terminal segments. **A**, Summary of V2 proteins predicted for ToCSV V30 and IR-swap mutants: V30ΔIR-s, V30ΔIR-sl, and V30ΔIR-sr infectious clones. Accepted (canonical) V2 regions in colour with aa length and identity indicated. Presence of putative N-terminal 30 aa segments (N-1 and N-2) shown using clear boxes. N-terminal segment **B**, PRALINE aa alignment sequence conservation; **C**, PSIPRED secondary structure (ss) prediction; and **D**, Phobius (top) and TMHMM v 2.0 (bottom) transmembrane helix predictions for N-1 and N2 aa sequences. Colour keys for each prediction shown.



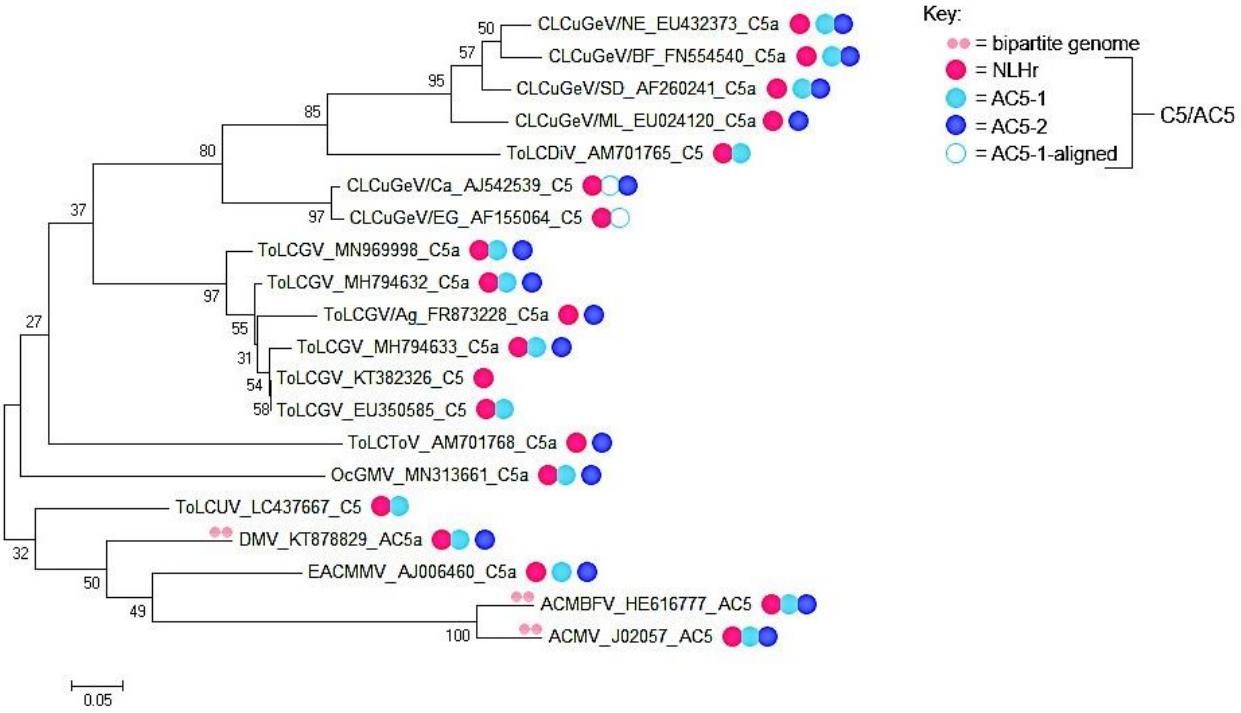
S2 Fig. 2.31 V2 multiple aa sequence alignment for ToCSV V30, V22, and selected African tomato-infecting begomoviruses sharing ≥78 % DNA-A sequence identity with ToCSV. **A**, PRALINE aa sequence alignment with colour key indicating aa conservation. **B**, PSIPRED secondary structure (ss) prediction with ss elements highlighted and consensus elements shown as cartoons. N-terminal region of interest indicated with black box. Details of all isolate sequences used in S2 Table 2.3.



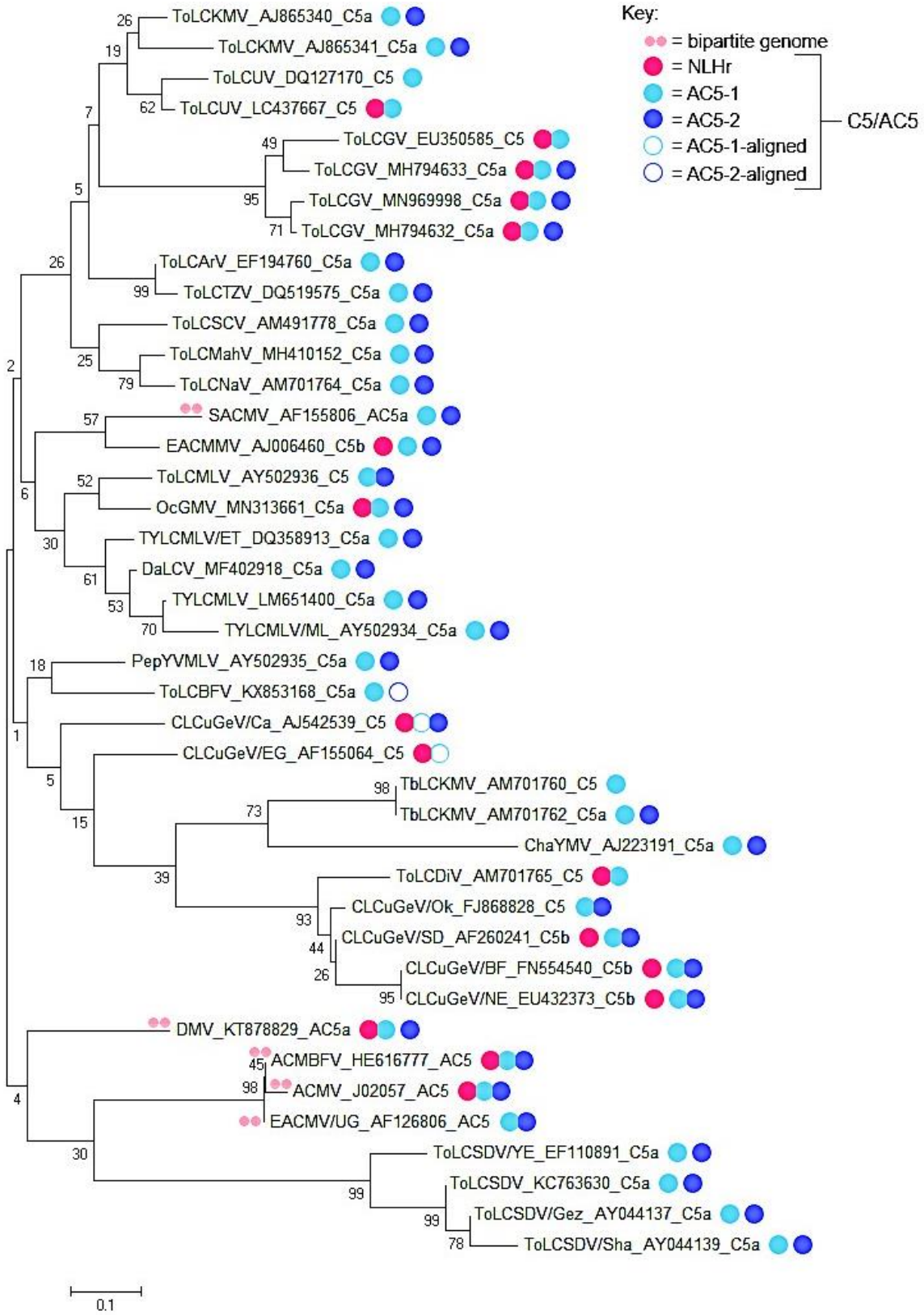
S2 Fig. 2.32 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30ΔV2-s, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



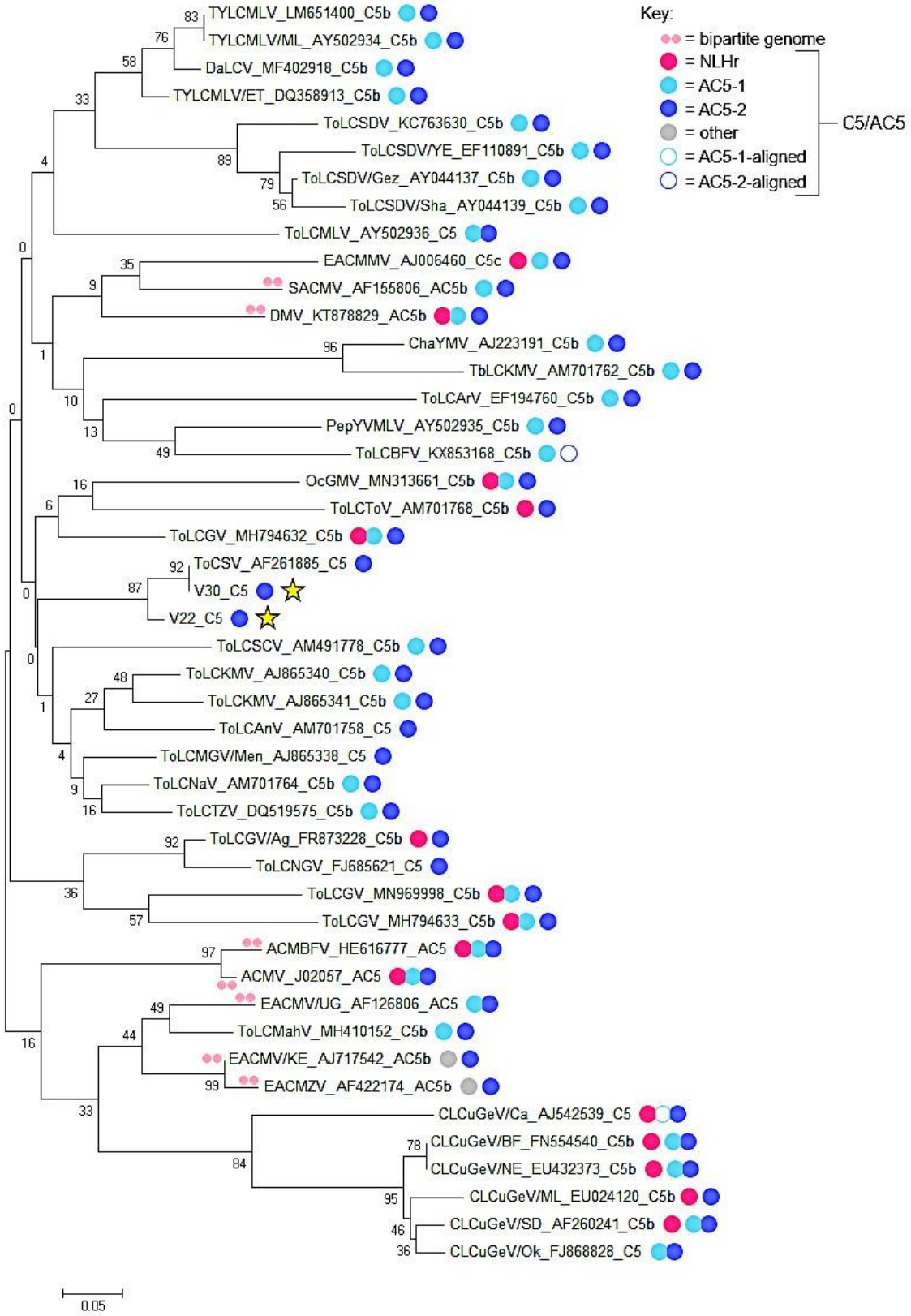
S2 Fig. 2.33 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V22ΔV2-s, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.



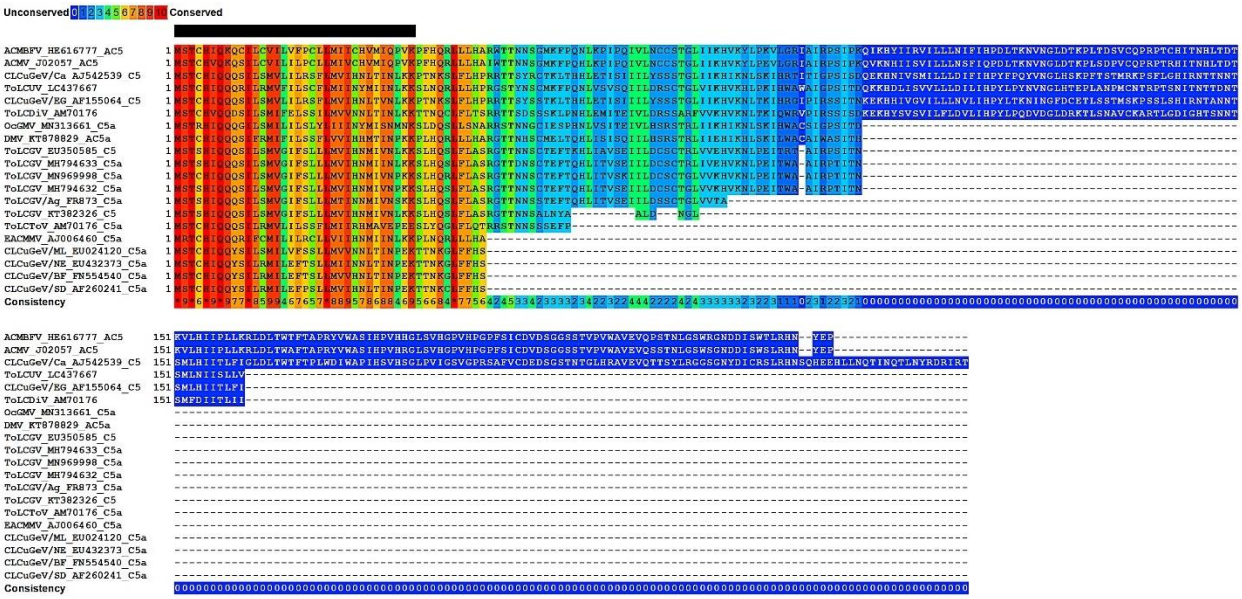
S2 Fig. 2.34 Phylogenetic tree constructed using full length C5/AC5 aa sequences of selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing N-terminal long helix region (NLHr). Alignment done using the neighbour-joining method (1,000 replications). Key indicates segmented translational profile of C5/AC5 ORF. Details of all isolate sequences used in S2 Table 2.3.



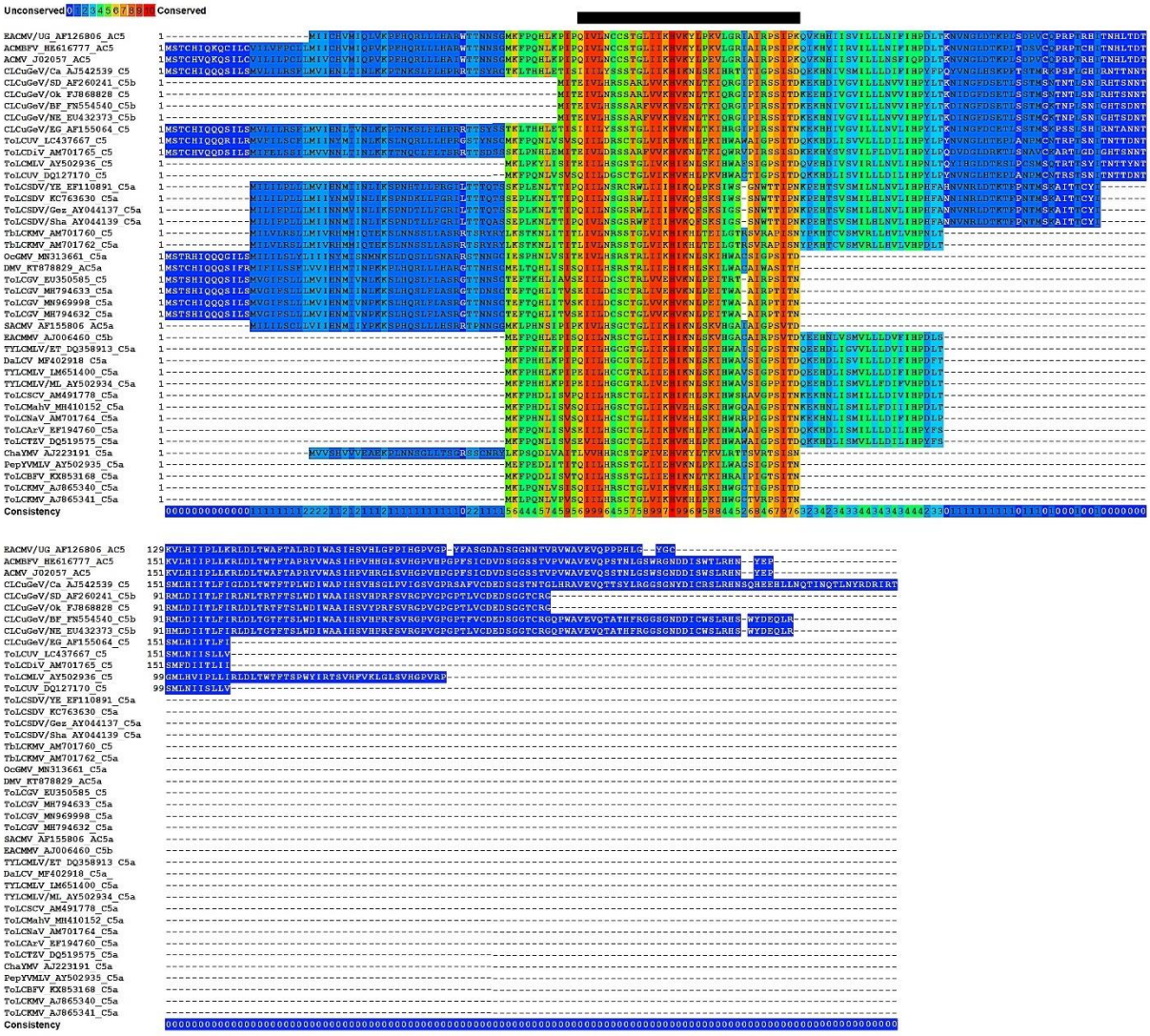
S2 Fig. 2.35 Phylogenetic tree constructed using full length C5/AC5 aa sequences of selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-1 domain. Alignment done using the neighbour-joining method (1,000 replications). Key indicates segmented translational profile of C5/AC5 ORF. Details of all isolate sequences used in S2 Table 2.3.



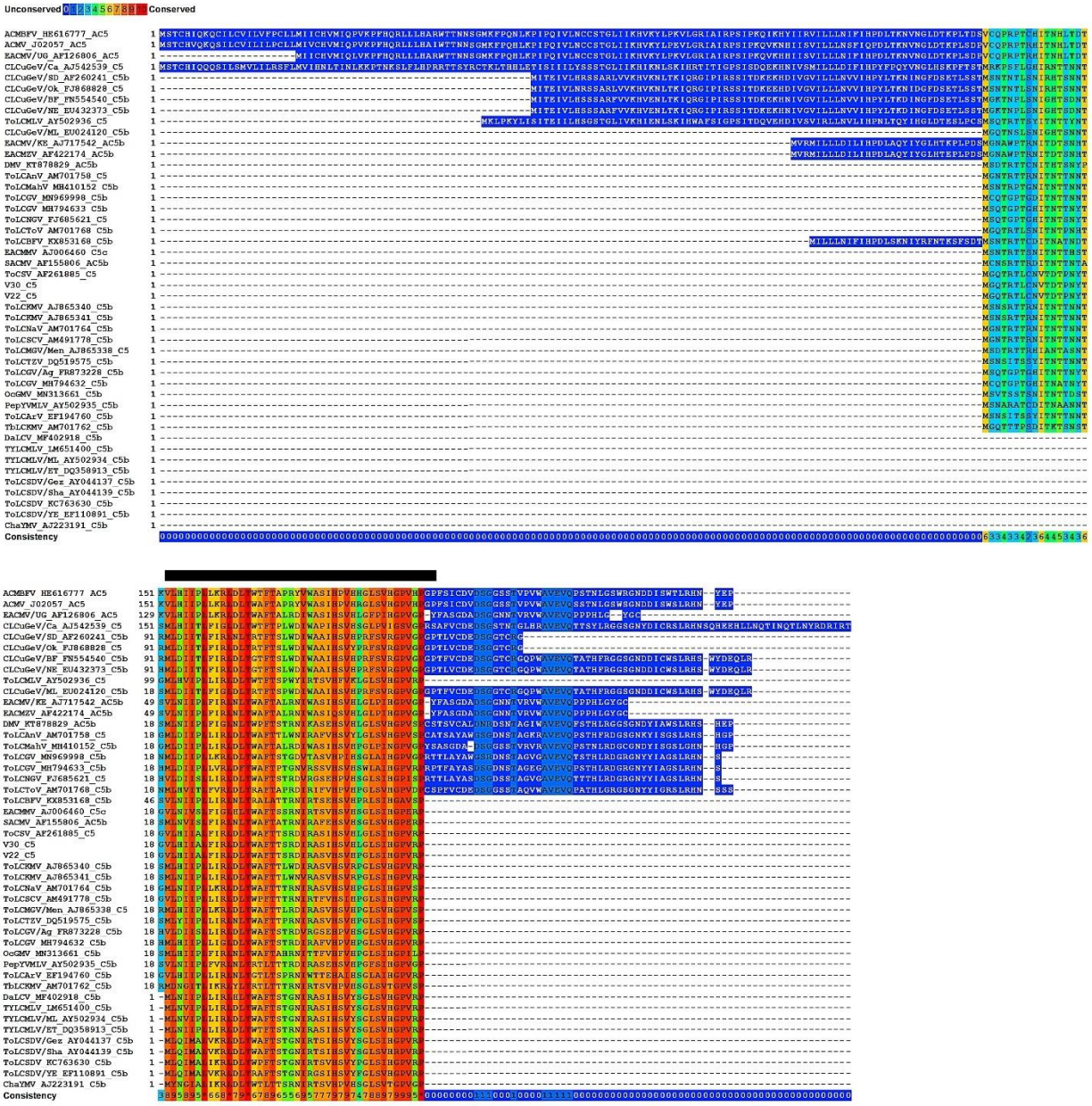
S2 Fig. 2.36 Phylogenetic tree constructed using full length C5/AC5 aa sequences of ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-2 domain. Alignment done using the neighbour-joining method (1,000 replications). Key indicates segmented translational profile of C5/AC5 ORF. V30 and V22 indicated using stars. Details of all isolate sequences used in S2 Table 2.3.



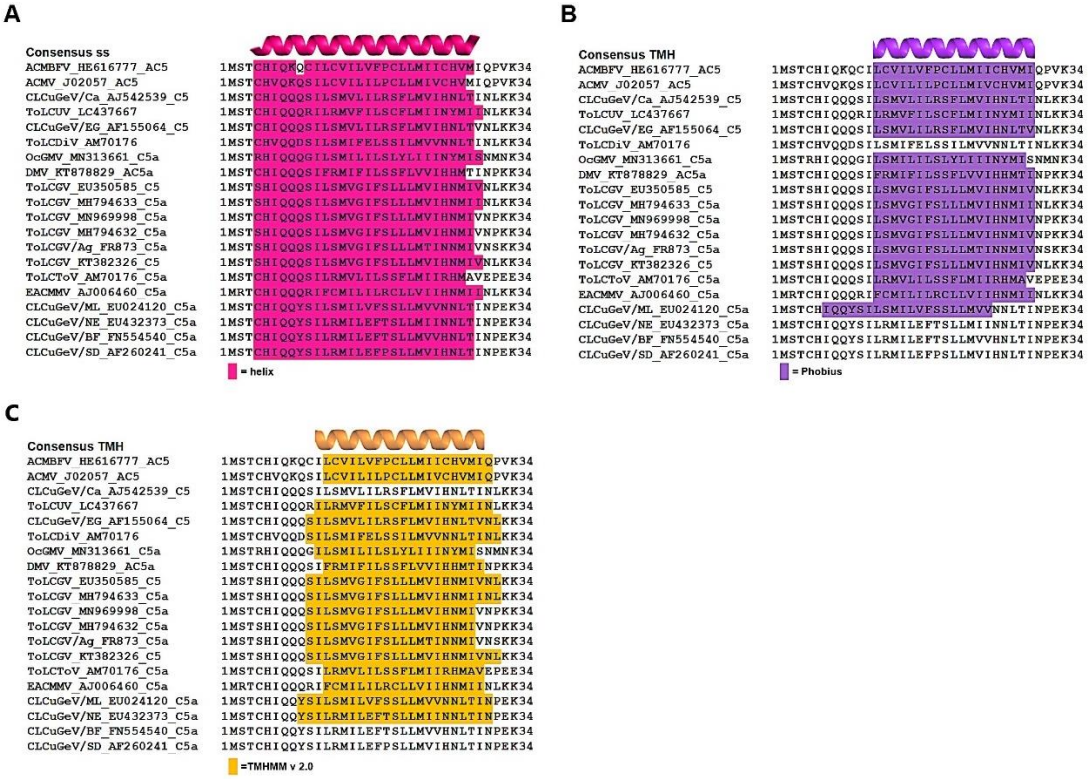
S2 Fig. 2.37 PRALINE C5/AC5 aa sequence alignment for selected African begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing N-terminal long helix region (NLHr). Colour key indicates aa conservation. NLHr indicated with black box. Details of all isolate sequences used in S2 Table 2.3.



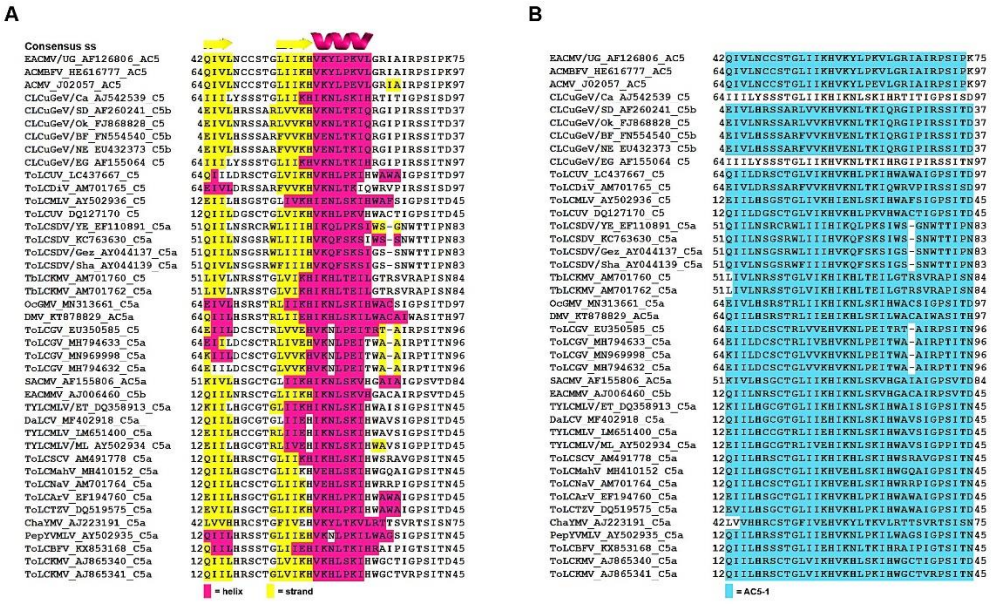
S2 Fig. 2.38 PRALINE C5/AC5 aa sequence alignment for selected African begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-1 domain. Colour key indicates aa conservation. AC5-1 indicated with black box. Details of all isolate sequences used in S2 Table 2.3.



S2 Fig. 2.39 PRALINE C5/AC5 aa sequence alignment for ToCSV V30, V22, and selected African begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-2 domain. Colour key indicates aa conservation. AC5-2 indicated with black box. Details of all isolate sequences used in S2 Table 2.3.

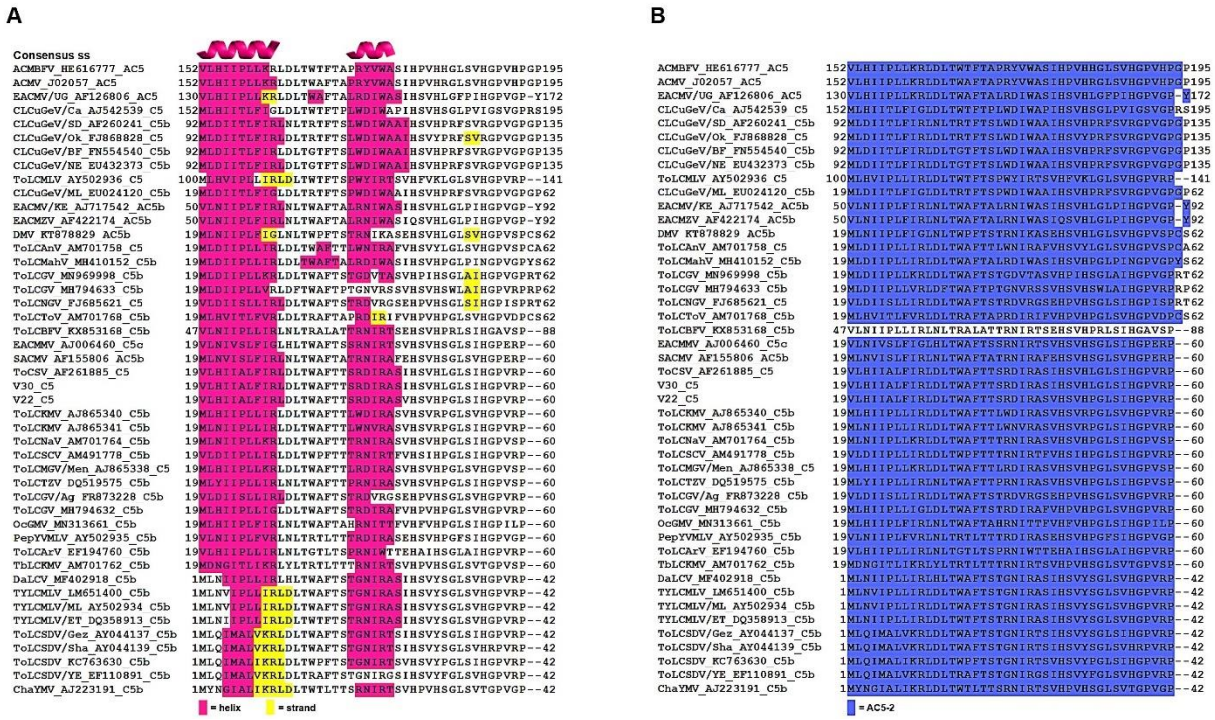


S2 Fig. 2.40 C5/AC5 N-terminal long helix region (NLHr) multiple aa sequence alignment. **A**, PSIPRED secondary structure; **B**, Phobius; **C**, TMHMM v 2.0 transmembrane helix predictions for selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing NLHr. Predicted elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.

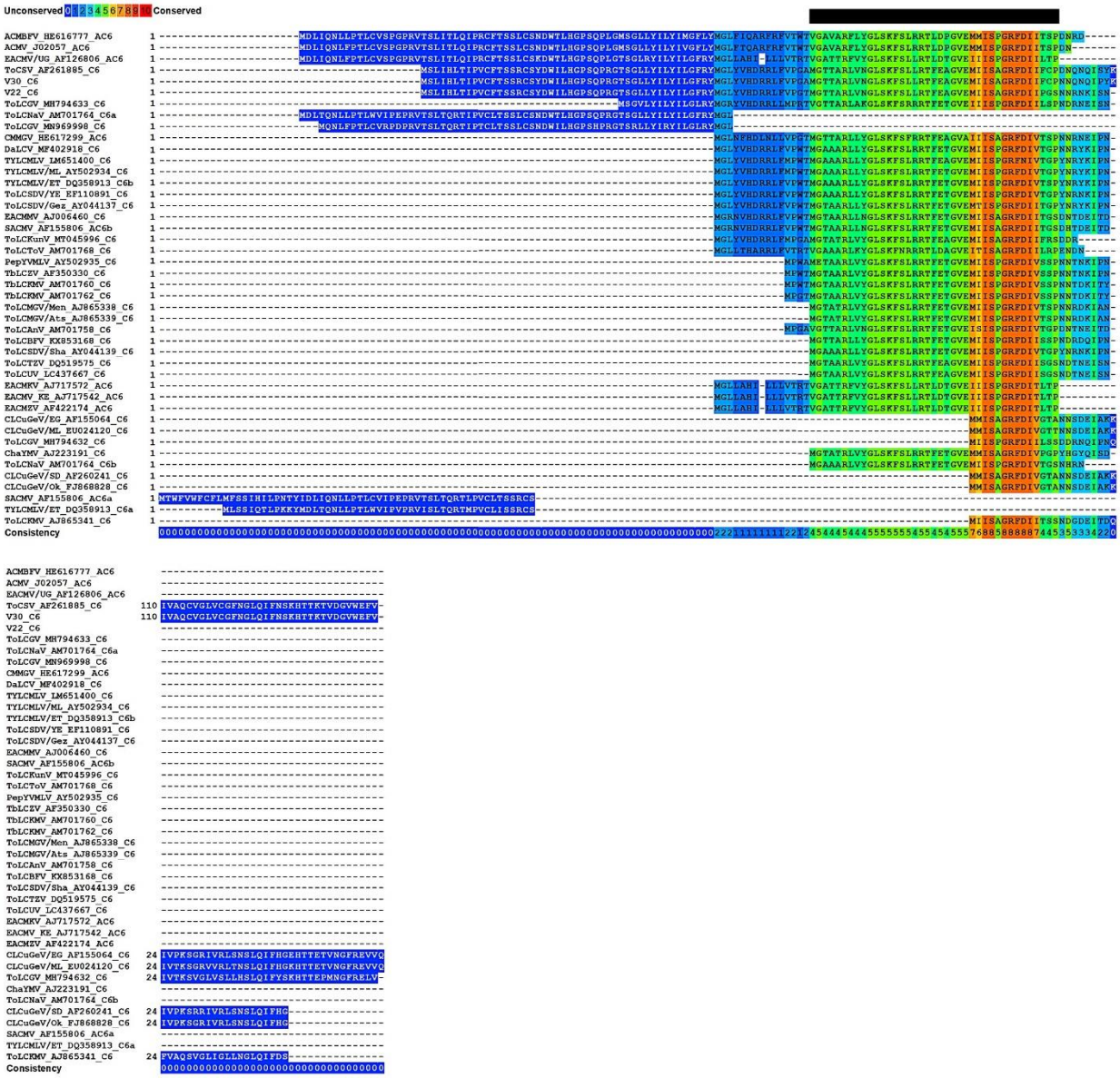


S2 Fig. 2.41 C5/AC5 AC5-1 domain multiple aa sequence alignment. **A**, PSIPRED secondary structure prediction; **B**, CDD domain identification for selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-1

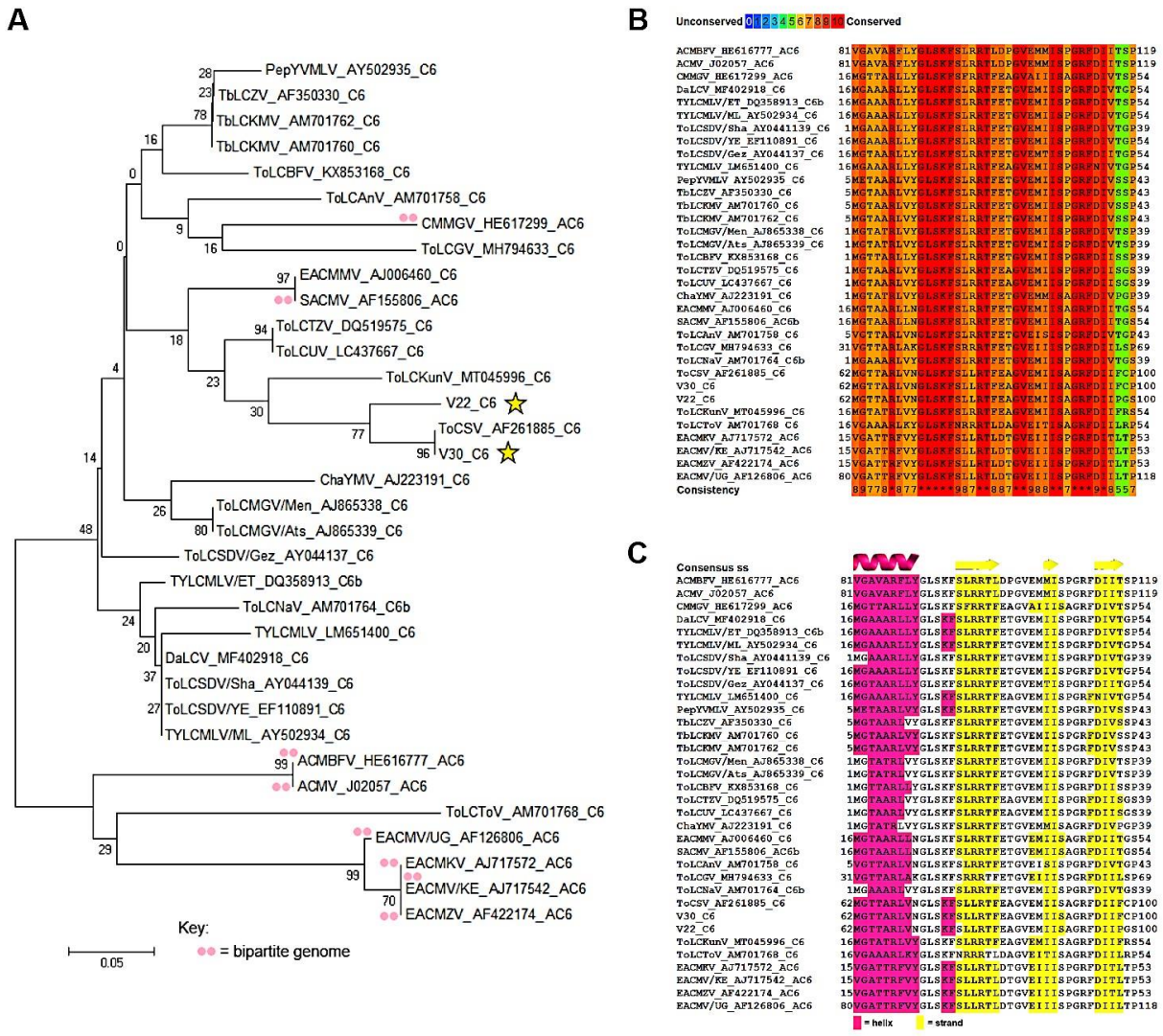
domain. Predicted ss elements highlighted and consensus elements shown as cartoons. AC5-1 (pfam04807) highlighted in cyan. Details of all isolate sequences used in S2 Table 2.3.



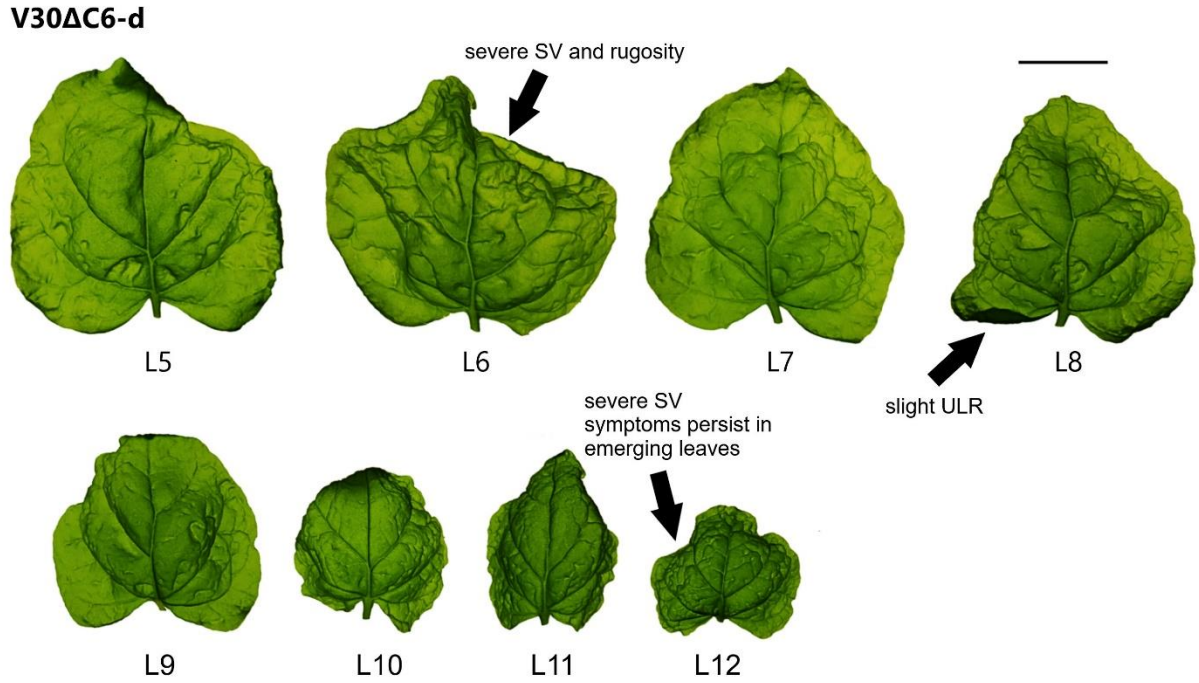
S2 Fig. 2.42 C5/AC5 AC5-2 domain multiple aa sequence alignment. **A**, PSIPRED secondary structure prediction; **B**, CDD domain identification for ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C5/AC5 proteins containing AC5-2 domain. Predicted ss elements highlighted and consensus elements shown as cartoons. AC5-2 (pfam08464) highlighted in navy. Details of all isolate sequences used in S2 Table 2.3.



S2 Fig. 2.43 PRALINE C6/AC6 aa sequence alignment for selected African begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV. Colour key indicates aa conservation. Putative conserved core domain (C6cd) indicated with black box. Details of all isolate sequences used in S2 Table 2.3.



S2 Fig. 2.44 Phylogenetic and aa sequence analysis of selected putative C6/AC6 proteins. **A**, Phylogenetic tree constructed using full length C6/AC6 aa sequences of ToCSV V30, V22, and selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C6/AC6 proteins containing a putative conserved core domain (C6cd). Alignment done using the neighbour-joining method (1,000 replications). V30 and V22 indicated using stars. **B**, PRALINE aa sequence alignment; **C**, PSIPRED secondary structure prediction for selected African mono- and bipartite begomoviruses sharing ≥75 % DNA-A sequence identity with ToCSV and having C6/AC6 proteins containing C6cd. Colour key indicates aa conservation. Predicted ss elements highlighted and consensus elements shown as cartoons. Details of all isolate sequences used in S2 Table 2.3.



S2 Fig. 2.45 Representative *Nicotiana benthamiana* plant inoculated with ToCSV V30ΔC6-d, dissected at 38 dpi. Leaf numbers indicate leaf emergence progression above uppermost agroinfiltrated leaf designated L0. Bar = 2 cm.

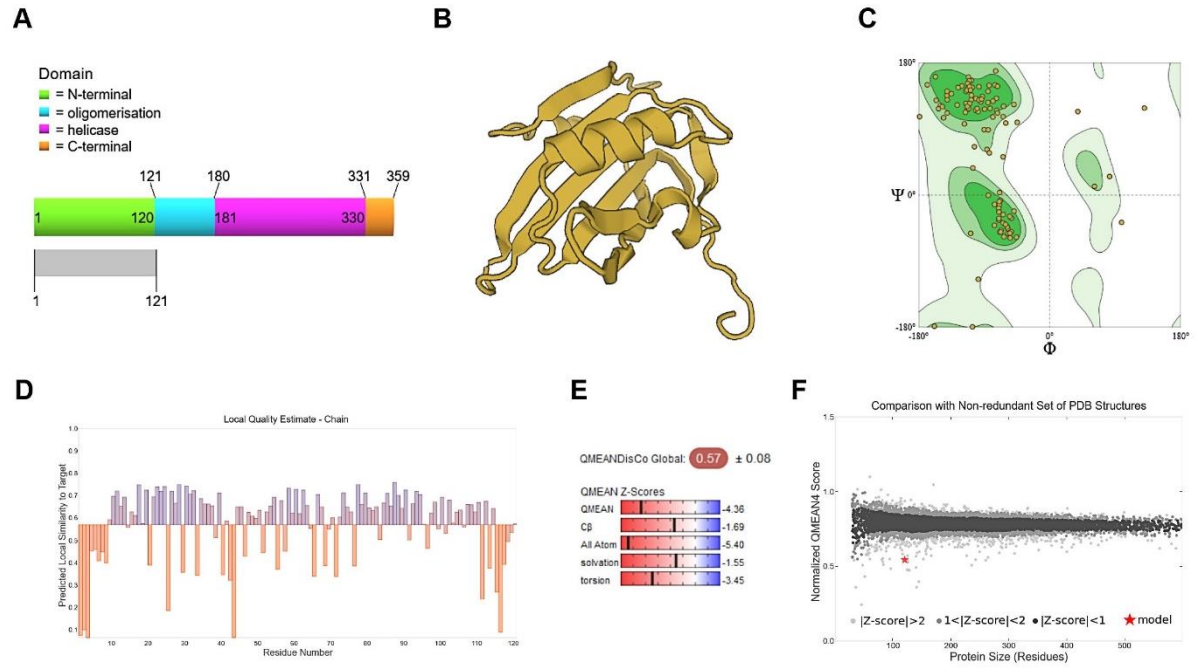
Supplementary 3 (S3) Figures

S3 Fig. 3.1 Sequence features of recombinant His-TF-RepC expressed using pCold TF-RepC. Nucleotide and translated aa sequence shown. Translation enhancing element in yellow, hexahistidine tag in blue, trigger factor chaperone tag in grey. Three cleavage sites indicated: HRV 3C protease in aquamarine, thrombin in pink, factor Xa in orange. A 10mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Codon-optimised ToCSV V30 RepC sequence was cloned using BamHI (green) and XbaI (violet).

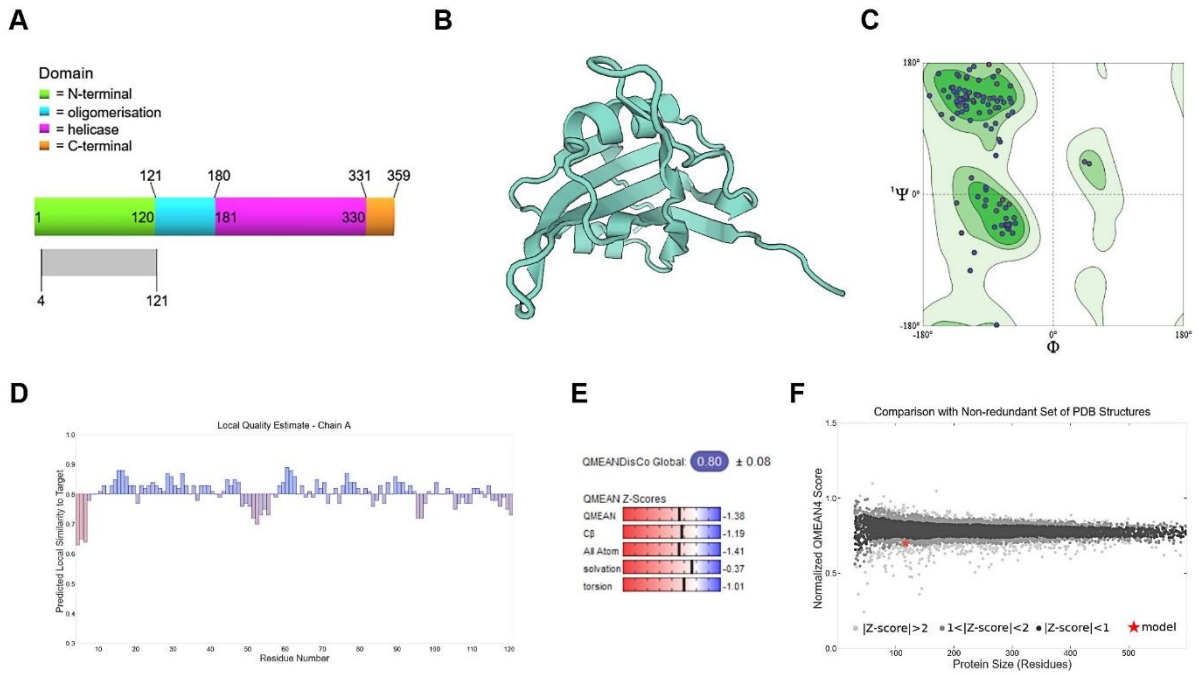
S3 Fig. 3.2 Sequence features of recombinant His-RepC expressed using pCold I-RepC. Nucleotide and translated aa sequence shown. Translation enhancing element in yellow, hexahistidine tag in blue. Factor Xa cleavage site in orange. A 10mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Codon-optimised ToCSV V30 RepC sequence was cloned using BamHI (green) and XbaI (violet).

```
atgaaaatcgaagaaggtaaactggtaatctggattaacggcgataaaggctataacggtctcgctgaagtcggtaagaaattcgagaaagataccggaattaaagtcaccgttgagcat
 M  K  I  E  E  G  K  L  V  I  W  I  N  G  D  K  G  Y  N  G  L  A  E  V  G  K  K  F  E  K  D  T  G  I  K  V  T  V  E  H
ccggataaactggaagagaaattcccacaggttgcggcaactggcgatggccctgacattatcttctgggcacacgaccgctttggtggctacgctcaatctggcctgttggctgaaatc
 P  D  K  L  E  E  K  F  P  Q  V  A  A  T  G  D  G  P  D  I  I  F  W  A  H  D  R  F  G  G  Y  A  Q  S  G  L  L  A  E  I
accccggacaaagcgttccaggacaagctgtatccgtttacctgggatgccgtacgttacaacggcaagctgattgcttacccgatcgctgttgaagcgttatcgctgatttataacaaa
 T  P  D  K  A  F  Q  D  K  L  Y  P  F  T  W  D  A  V  R  Y  N  G  K  L  I  A  Y  P  I  A  V  E  A  L  S  L  I  Y  N  K
gatctgctgccgaacccgccaaaaacctgggaagagatcccggcgctggataaagaactgaaagcgaaaggtaagagcgcgctgatgttcaacctgcaagaaccgtacttcacctggccg
 D  L  L  P  N  P  P  K  T  W  E  E  I  P  A  L  D  K  E  L  K  A  K  G  K  S  A  L  M  F  N  L  Q  E  P  Y  F  T  W  P
ctgattgctgctgacgggggttatgcgttcaagtatgaaaacggcaagtacgacattaaagacgtgggcgtggataacgctggcgcgaaagcgggtctgaccttcctggttgacctgatt
 L  I  A  A  D  G  G  Y  A  F  K  Y  E  N  G  K  Y  D  I  K  D  V  G  V  D  N  A  G  A  K  A  G  L  T  F  L  V  D  L  I
aaaaacaaacacatgaatgcagacaccgattactccatcgcagaagctgcctttaataaaggcgaaacagcgatgaccatcaacggcccgtgggcatggtccaacatcgacaccagcaaa
 K  N  K  H  M  N  A  D  T  D  Y  S  I  A  E  A  A  F  N  K  G  E  T  A  M  T  I  N  G  P  W  A  W  S  N  I  D  T  S  K
gtgaattatggtgtaacggtactgccgaccttcaagggtcaaccatccaaaccgttcgttggcgtgctgagcgcaggtattaacgccgccagtccgaacaaagagctggcaaaagagttc
 V  N  Y  G  V  T  V  L  P  T  F  K  G  Q  P  S  K  P  F  V  G  V  L  S  A  G  I  N  A  A  S  P  N  K  E  L  A  K  E  F
ctcgaaaactatctgctgactgatgaaggtctggaagcggttaataaagacaaaccgctgggtgccgtagcgctgaagtcttacgaggaagagttggcgaaagatccacgtattgccgcc
 L  E  N  Y  L  L  T  D  E  G  L  E  A  V  N  K  D  K  P  L  G  A  V  A  L  K  S  Y  E  E  E  L  A  K  D  P  R  I  A  A
actatggaaaacgcccagaaaggtgaaatcatgccgaacatcccgcagatgtccgctttctggtatgccgtgcgtactgcggtgatcaacgccgccagcggtcgtcagactgtcgatgaa
 T  M  E  N  A  Q  K  G  E  I  M  P  N  I  P  Q  M  S  A  F  W  Y  A  V  R  T  A  V  I  N  A  A  S  G  R  Q  T  V  D  E
gccctgaaagacgcgcagactaattcgagctcgaacaacaacaacaatacaatacaacaacctcgggatcgagggaaggatttcagaattcggatccagcgcgcgtggtggctgccag
 A  L  K  D  A  Q  T  N  S  S  S  N  N  N  N  N  N  N  N  N  L  G  I  E  G  R  I  S  E  F  G  S  S  A  R  G  G  C  Q
aacgcgaacgatgcgtgcgcggatgcgctgaacgcggacagcgcggaggcggcgatggcgatcattcgtgaaaagctgccgaaagacttcatctttcagtaccacaacctgaaaaccaac
 N  A  N  D  A  C  A  D  A  L  N  A  D  S  A  E  A  A  M  A  I  I  R  E  K  L  P  K  D  F  I  F  Q  Y  H  N  L  K  T  N
ctggatcgtatcttccaagctccgccggagccgtatattagcccgttcagcagcagcagctttgaccaagtgccggaggaactggaggaatgggttagcgagaacgtttgcggtgcggcg
 L  D  R  I  F  Q  A  P  P  E  P  Y  I  S  P  F  S  S  S  S  F  D  Q  V  P  E  E  L  E  E  W  V  S  E  N  V  C  G  A  A
gcgcgtccgtggcgtccgaacagcatcgttattgaaggtgatagccgtagcggcaaaaccatctgggcgcgtagcctgggtccgcacaactacctgtgcggccacctggacctgagcccg
 A  R  P  W  R  P  N  S  I  V  I  E  G  D  S  R  S  G  K  T  I  W  A  R  S  L  G  P  H  N  Y  L  C  G  H  L  D  L  S  P
aaggtgtacagcaacgatgcgtggtataacgtgattgacgatgttgacccgcactacctgaagcactggaaagaattctttggtgcgcagcgtgattggcaaagcaacaccaagtatggc
 K  V  Y  S  N  D  A  W  Y  N  V  I  D  D  V  D  P  H  Y  L  K  H  W  K  E  F  F  G  A  Q  R  D  W  Q  S  N  T  K  Y  G
aaaccgatccagattaaaggtggcatcccgaccattttcctgtgcaacccgggtccgaccagcagctacaaggagtatctggacgaagataagaacaccagcctgaaagcgtggaccatc
 K  P  I  Q  I  K  G  G  I  P  T  I  F  L  C  N  P  G  P  T  S  S  Y  K  E  Y  L  D  E  D  K  N  T  S  L  K  A  W  T  I
aagaacgcggttttattaccctgaaccagccgctgtatagcagcccgaaccaaggtccggcgagcaacagccaggaagagagcagccaagcgtaatctaga
 K  N  A  V  F  I  T  L  N  Q  P  L  Y  S  S  P  N  Q  G  P  A  S  N  S  Q  E  S  S  Q  A  –
```

S3 Fig. 3.3 Sequence features of recombinant MBP-RepC expressed using pMAL-c2X-RepC. Nucleotide and translated aa sequence shown. Maltose binding protein tag in grey. Factor Xa cleavage site in orange. A 6mer linker is located upstream of RepC (red) which includes the helicase domain (Rep aa 122-330) and the C-terminal domain (Rep aa 331-359). Codon-optimised ToCSV V30 RepC sequence was cloned using BamHI (green) and XbaI (violet).



S3 Fig. 3.4 SWISS-MODEL structure quality assessment of ToCSV V30 Rep N-terminal domain homology model monomer generated using Phyre2. **A**, Map of V30 Rep domain positions with coordinates of sequence modelled indicated in grey. Homology model **B**, cartoon representation; **C**, Ramachandran plot; **D**, local quality estimate; **E**, QMEANDisCo Global and QMEAN Z-scores; **F**, Z-score comparison with non-redundant set of PDB structures.



S3 Fig. 3.5 SWISS-MODEL structure quality assessment of ToCSV V30 Rep N-terminal domain homology model monomer generated using SWISS-MODEL. **A**, Map of V30 Rep domain positions with coordinates of sequence modelled indicated in grey. Homology model **B**, cartoon representation; **C**, Ramachandran plot; **D**, local quality estimate; **E**, QMEANDisCo Global and QMEAN Z-scores; **F**, Z-score comparison with non-redundant set of PDB structures.

>V30_Rep_Phyre2|Chain A|
MAPPKRFQINSKNYFVTYPKCSLTKEESLSQLLNLETPTNKKYIKICRELHENGEPHLHVLIQFEGKFNCKNHRF
FDLVSPNRSAHFHPNIQGAKSSSDVKSYIDKDGDTLEWGEFQVDGR

>V30_Rep_SWISS-MODEL|Chain A|
PKRFQINSKNYFVTYPKCSLTKEESLSQLLNLETPTNKKYIKICRELHENGEPHLHVLIQFEGKFNCKNHRFFDL
VSPNRSAHFHPNIQGAKSSSDVKSYIDKDGDTLEWGEFQVDGR

>1L5I_1|Chain A|Rep protein|Tomato yellow leaf curl Sardinia virus (123735)
SGRFSIKAKNYFLTYPKCDLTKENALSQITNLQTPTNKLFIKICRELHENGEPHLHILIQFEGKYNCTNQRFFDL
VSPTRSAHFHPNIQGAKSSSDVKSYIDKDGDVLEWGTFQIDGR

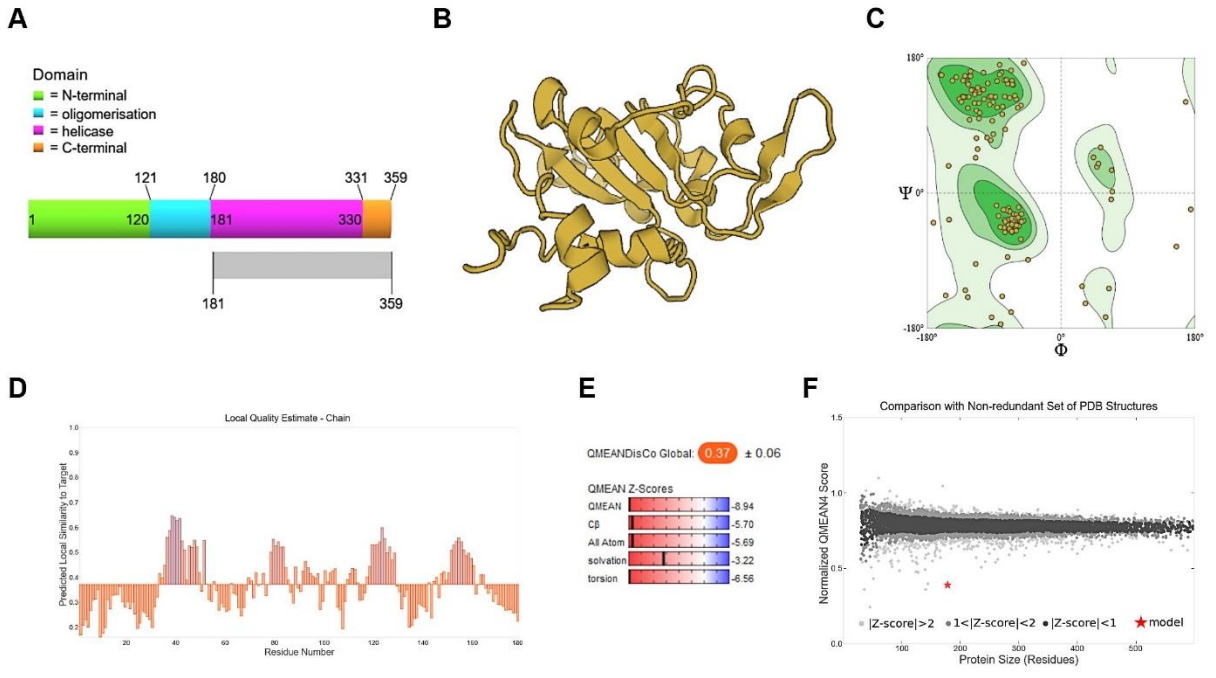
>6Q1M_1|Chain A|Replication-associated protein|Wheat dwarf virus (10834)
FRVYSKYLFLTYPQCTLEPQYALDSLRTLLNKYEPLYIAAVRELHEDGSPHLHVLVQNKLRASITNPNALNLRMD
TSPFSIFHPNIQAAKDCNQVRDYITKEVDSDVNTAEWGTFVAVS

>2HWT_1|Chain A|Putative replicase-associated protein|Faba bean necrotic
yellows virus (59817)
ARQVICWCFTLNNPLSPLSLHDSMKYLVYQTEQGEAGNIHFQGYIEMKKRTSLAGMKKLIPGAHFEKRRGTQGEA
RAYSMKEDTRLEGPWEYGE

>2HW0_1|Chain A|Replicase|Porcine circovirus 2 (85708)
PSKKNGRSGPQPHKRWVFTLNNPSEDERKKIRDLPISLFDYFIVGEEGNEEGRTPHLQGFANFVKKQTFNKVKWY
LGARCHIEKAKGTDQQNKEYCSKEGNLLMECGAPRSQGQR

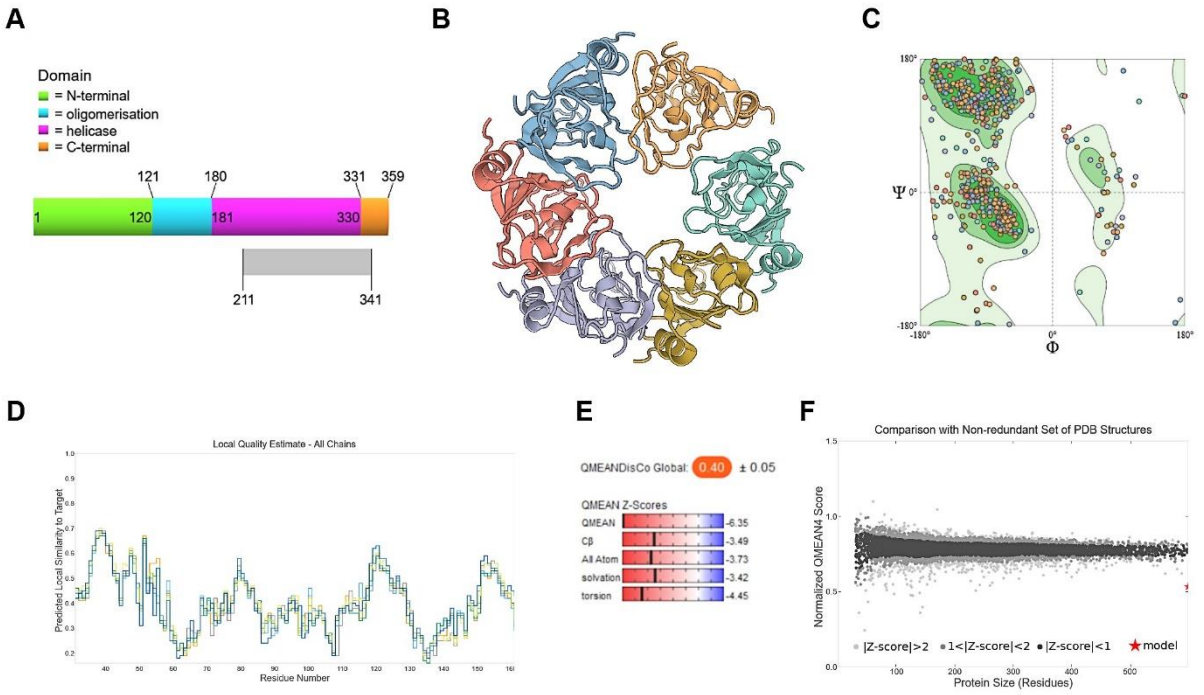
S3 Fig. 3.6 PyMOL-assigned secondary structure (ss) for ToCSV V30 Rep N-terminal domain homology model monomers generated using Phyre2 and SWISS-MODEL; and solved monomeric structures of TYLCSV Rep (PDB ID: 1L5I), WDV Rep (PDB ID: 6Q1M), FBNYV M-Rep (PDB ID: 2HWT), and PCV2 Rep (PDB ID: 2HW0). Helices highlighted in magenta, strands in yellow.

S3 Fig. 3.7 Rep oligomerisation domain multiple sequence alignment for ToCSV V30, V22, TYLCSV (X61153), and selected African mono- and bipartite begomoviruses. MUSCLE aa sequence alignment with 100 % and 70 % aa sequence consensus shown. Consensus residues capitalised and coloured lowercase letter key indicates consensus residue type. PSIPRED secondary structure (ss) and DISOPRED protein disorder prediction, with ss elements highlighted and disordered residues coloured according to colour key. Consensus ss elements at confidence (conf.) levels ≥0 and ≥5 indicated. Rep sequence obtained with ORF finder using the DNA-A sequence as the input. DNA-A sequences for all isolates used in S3 Table 3.2.



S3 Fig. 3.8 SWISS-MODEL structure quality assessment of ToCSV V30 Rep helicase and C-terminal domain homology model monomer generated using Phyre2. **A**, Map of V30 Rep domain positions with coordinates of sequence modelled indicated in grey. Homology model **B**, cartoon representation; **C**,

Ramachandran plot; **D**, local quality estimate; **E**, QMEANDisCo Global and QMEAN Z-scores; **F**, Z-score comparison with non-redundant set of PDB structures.

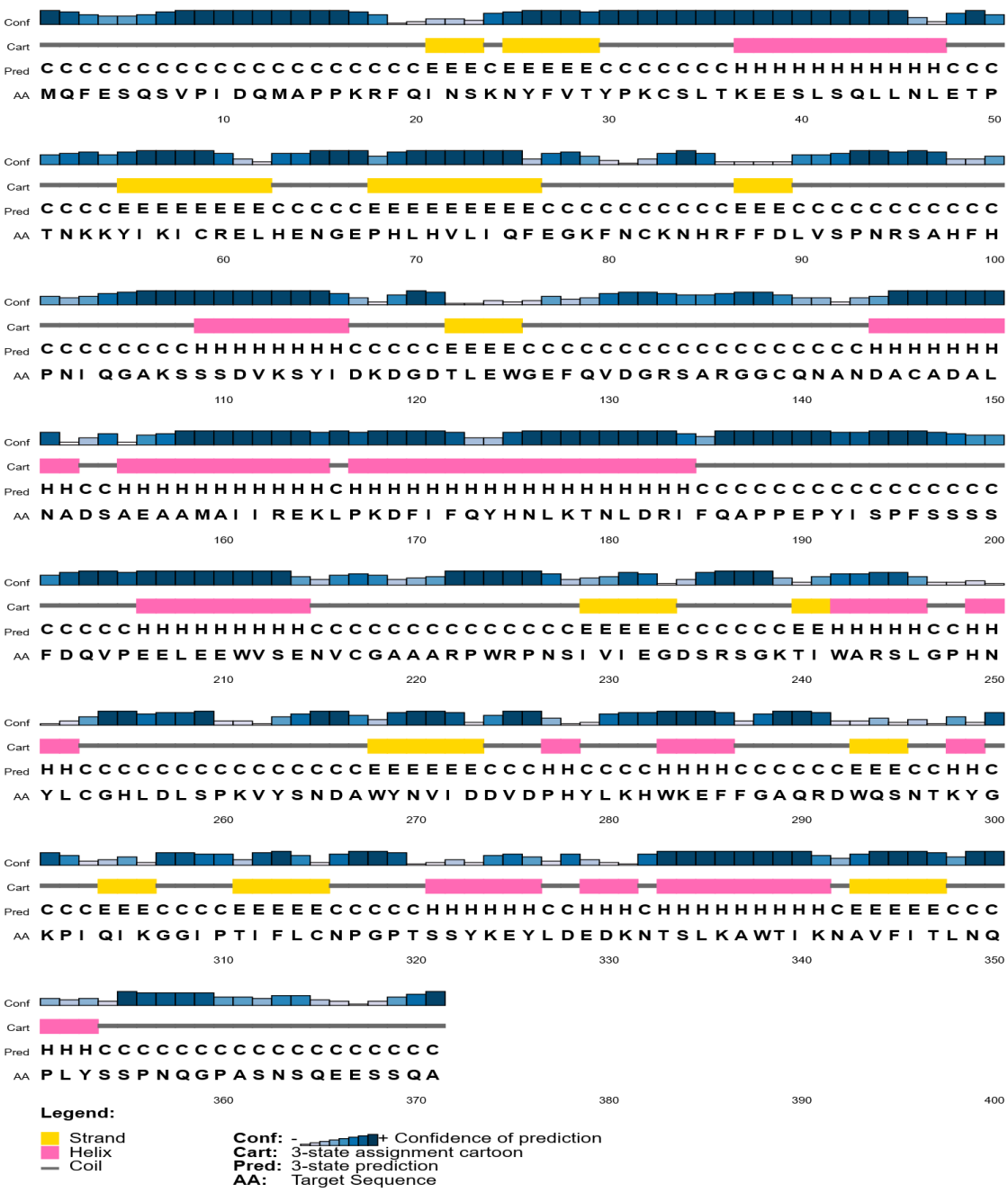


S3 Fig. 3.9 SWISS-MODEL structure quality assessment of ToCSV V30 Rep partial helicase and C-terminal domain homology model hexamer generated using SWISS-MODEL. **A**, Map of V30 Rep domain positions with coordinates of sequence modelled indicated in grey. Homology model **B**, cartoon representation; **C**, Ramachandran plot; **D**, local quality estimate; **E**, QMEANDisCo Global and QMEAN Z-scores; **F**, Z-score comparison with non-redundant set of PDB structures.

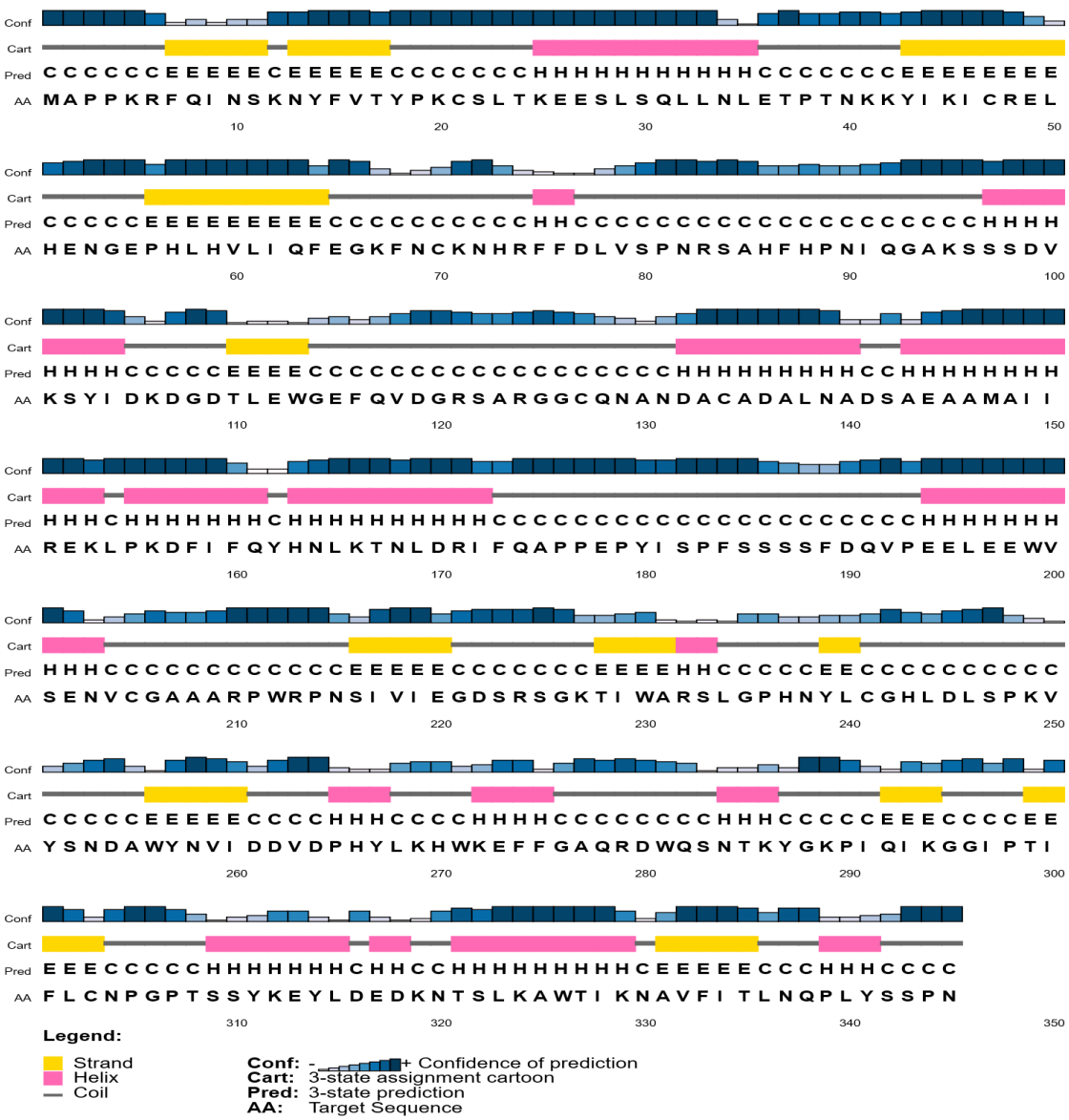
```
>V30_RepPhyre2|Chain A|
ISPFSSSSFDQVPEELEEWVSENVCGAAARPWRPNSIVIEGDSRSGKTIWARSLGPHNYLCGHLDLSPKVYSNDA
WYNVIDDVDPHYLKHWKEFFGAQRDWQSNTKYGKPIQIKGGIPTIFLCNPGPTSSYKEYLDEDKNTSLKAWTIKN
AVFITLNQPLYSSPNQGPASNSQEESSQA


>V30_Rep_SWISS-MODEL|Chains A, B, C, D, E, F| (consensus ss)
PWRPNSIVIEGDSRSGKTIWARSLGPHNYLCGHLDLSPKVYSNDAWYNVIDDVDPHYLKHWKEFFGAQRDWQSNT
KYGKPIQIKGGIPTIFLCNPGPTSSYKEYLDEDKNTSLKAWTIKNAVFITLNQPLY

>7JSI_1|Chains A, B, C, D, E, F|Protein Rep68|Adeno-associated virus - 2
(10804) (partial sequence, consensus ss)
FGKRNTIWLFGPATTGKTNIAEAIAHTVPFYGCVNWTNENFPFNDCVDKMVIWWEEGKMTAKVVESAKAILGGSK
VRVDQKCKSSAQIDPTPVIVTSNTNMCAVIDGNSTTFEHQQPLQDRMFKFELTRRL

>7LAR_1|Chains A, B, C, D, E, F|ATP-dependent helicase Rep|Porcine
circovirus 2 (85708) (partial sequence, consensus ss)
DWKTNVHVIVGPPGCGKSKWAANFADPETTYWKPPRNKWWDGYHGEEVVVIDDFYGWLPWDDLLRLCDRYPLTVE
TKGGTVPFLARSILITSNQTPLEWYSSTAVPAVEALYRRITSLVFWKNATEQSTEEGG

>2GXA_2|Chains C[auth A], D[auth B], E[auth C], F[auth D], G[auth E],
H[auth F], I[auth G], J[auth H], K[auth I], L[auth J], M[auth K], N[auth
L]|Replication protein E1|Bovine papillomavirus type 1 (337052) (partial
sequence, consensus ss)
IPKKNCLAFIGPPNTGKSMLCNSLIHFLGGSVLSFANHKSHFWLASLADTRAALVDDATHACWRYFDTYLRNALD
GYPVSIDRKHKAAVQIKAPPLLVTSNIDVQAEDRYLYLHSRVQTFRFEQPCTDESG

>1SVM_1|Chains A, B, C, D, E, F|large T antigen|Simian virus 40 (1891767)
(partial sequence, consensus ss)
IPKKRYWLFKGPIDSGKTTLAAALLELCGGKALNVNLPLDRLNFELGVAIDQFLVVFEDVKGTGGESRDLPSGQG
INNLDNLRDYLDGSVKVNLEKKHLNKRTQIFPPGIVTMNEYSVPKTLQARFVKQIDFRP
```

S3 Fig. 3.10 PyMOL-assigned secondary structure (ss) for ToCSV V30 Rep full helicase and C-terminal domain monomer homology model generated using Phyre2, and partial V30 Rep helicase and C-terminal domain hexamer homology model generated using SWISS-MODEL; as well as partial solved hexameric structures of AAV2 Rep (PDB ID: 7JSI), PCV2 Rep (PDB ID: 7LAR), BPV E1 (PDB ID: 2GXA), and SV40 Tag (PDB ID 1SVM). For all hexameric structures, consensus monomer ss is shown. Helices highlighted in magenta, strands in yellow.

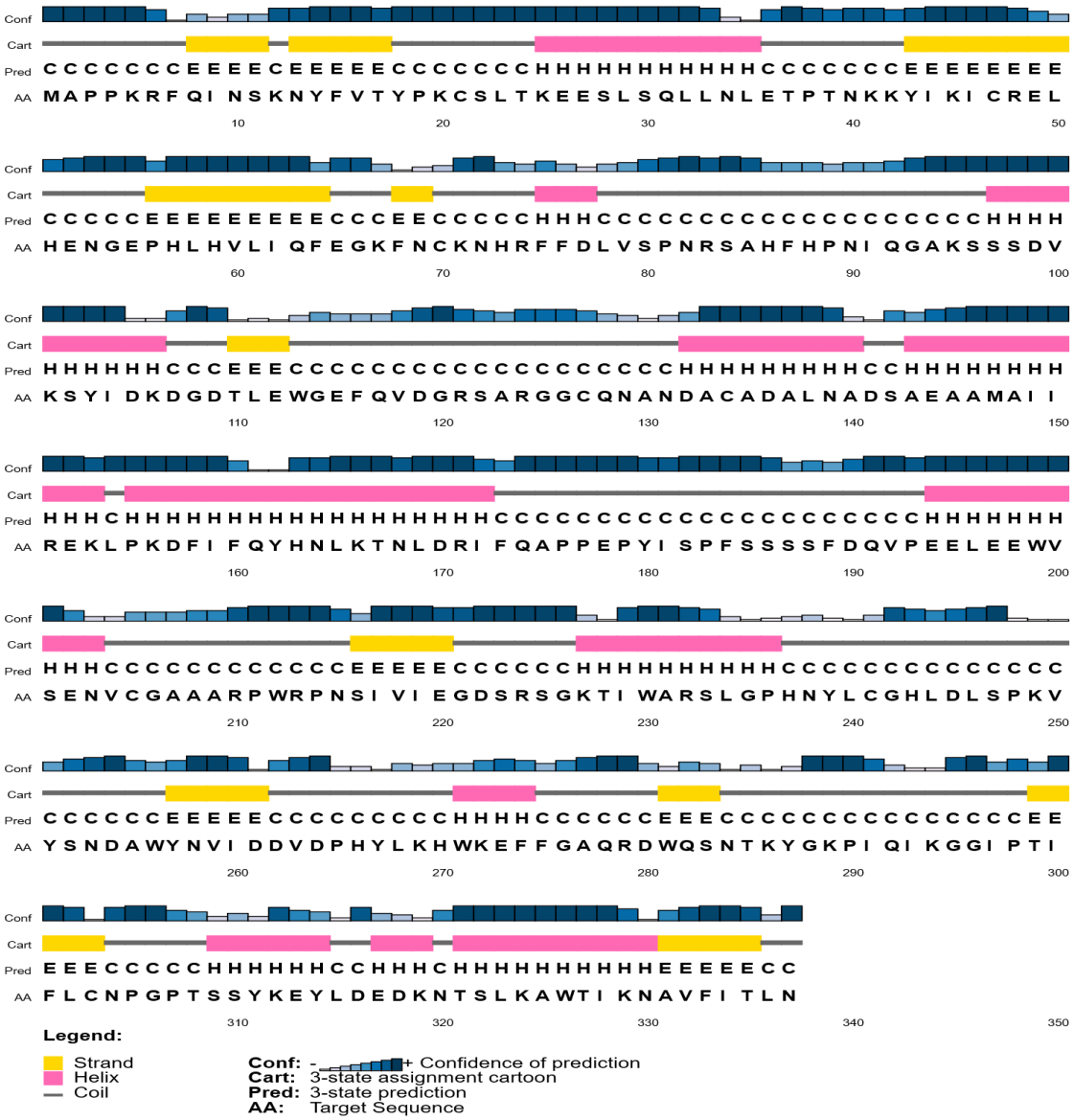
S3 Fig. 3.11 PSIPRED secondary structure (ss) prediction for ToCSV V30 canonical Rep. Amino acid position 1 is canonical methionine. Three state ss prediction with cartoon and prediction confidence levels indicated.



S3 Fig. 3.12 PSIPRED secondary structure (ss) prediction for ToCSV V30 Rep containing putative 12 aa N-terminal extension (Rep-N). Amino acid position 13 is canonical methionine. Three state ss prediction with cartoon and prediction confidence levels indicated.



S3 Fig. 3.13 PSIPRED secondary structure (ss) prediction for ToCSV V30ΔC1-d(14) Rep with canonical start methionine. Three state ss prediction with cartoon and prediction confidence levels indicated.



S3 Fig. 3.14 PSIPRED secondary structure (ss) prediction for ToCSV V30ΔC1-d(22) Rep with canonical start methionine. Three state ss prediction with cartoon and prediction confidence levels indicated.

Supplementary 2 (S2) Tables

S2 Table 2.1 Description of sequence changes made to generate ToCSV variant IR and V2 sequence swap mutants.

Mutant	Template	Sequence removed	Sequence inserted
V30ΔIR-s	V30	V30 nt 61 to 138 (partial IR 3' region)	V22 nt 61 to 137 (partial IR 3' region)
V30ΔIR-sl	V30	V30 nt 61 to 94 (partial IR 3' region to left of TATA box)	V22 nt 61 to 94 (partial IR 3' region to left of TATA box)
V30ΔIR-sr	V30	V30 nt 108 to 138 (partial IR 3' region to right of TATA box)	V22 nt 108 to 137 (partial IR 3' region to right of TATA box)
V30ΔV2-s	V30	V30 nt 139 to 489 (full canonical V2 ORF region)	V22 nt 138 to 488 (full canonical V2 ORF region)
V22ΔV2-s	V22	V22 nt 138 to 488 (full canonical V2 ORF region)	V30 nt 139 to 489 (full canonical V2 ORF region)

S2 Table 2.2 Primers used in this study.

Primer	Sequence (5'-3')	Purpose
ToCSVseq_FP	GTCGTGCTCCACCATGTTG	PCR, sequencing (IR)
ToCSVseq_RP	CACCACGAACCTTTTACCCG	PCR, sequencing (IR)
V22IR_FP2	AAACTCCTAAAGCGGCCATC	PCR, sequencing (IR)
V22V2_FP	GTGCGGATGTCATAATGTGG	PCR, sequencing (V2)
V22V2_RP	GACCAGGACCCTTTGTAATGTC	PCR, sequencing (V2)
ΔC6del_FP	GCCACGATCTaATACGGGATCTGATTCTGG	Mutagenesis to generate V30ΔC6-d
ΔC6del_RP	CCCGTATTtAGATCGTGGCCCAATGTGTTG	
C5seq_FP	TGGGATCCACTGTTAAACGA	cDNA synthesis, PCR (for RT-PCR), sequencing (C6)
C5seq_RP	TCACCACGAACCTTTTACCC	cDNA synthesis, PCR (for RT-PCR), sequencing (C6)
GAPDH_FP	CTGTTATTGGAGGAGGGAACAA	qPCR
GAPDH_RP	AGTCTTTCCTACCATGCCAAC	qPCR
V2_FP	CGACAGCCCGTTCACCAG	qPCR
V2_RP	GCCTGTACGTCCATGATCGTC	qPCR

V30IR_FP	GCATTTCAAAATTCAAACTCCTAAAGC	cDNA synthesis, PCR (for RT-PCR)
V30IR_RP	AGCAGATAACTAAGCGATAAGGC	cDNA synthesis, PCR (for RT-PCR)
V30V2N_FP	CCAATCAGAATGTCGCCTTA	PCR (for RT-PCR)
V30V2N_RP	CAATGTGTTGGGCTCGTATG	cDNA synthesis, PCR (for RT-PCR)
V22V2N_FP	TGAAATTGAGGCGTTAAAGCT	PCR (for RT-PCR)
V22V2N_RP	ACGAAACCCATGAACAGAGTCT	cDNA synthesis, PCR (for RT-PCR)
V30C6_FP	TCCACGGTTTTCGTTGTATG	cDNA synthesis, PCR (for RT-PCR)
C6_RP	TGGATCTTACATGGGCCTTC	PCR
V30M_FP	GAAACGACCAGTCTGAGGCT	cDNA synthesis, PCR (for RT-PCR)
V30M_RP	TAAAGGCGGCATTCCCACTA	cDNA synthesis, PCR (for RT-PCR)
C5M_FP	ACGATCATGGACGTACAGGC	cDNA synthesis, PCR (for RT-PCR)
C1T_RP	CCCTCAACCAGCCACTGTAC	cDNA synthesis, PCR (for RT-PCR)

Key: lower case, target nt mutation.

S2 Table 2.3 African mono- and bipartite begomovirus isolates used in this study.

Begomovirus species	**Abbreviation**	**Isolate**	**DNA-A accession number**	**DNA-A pairwise sequence identity (%)**	
				ToCSV-[ZA:Mks30:08]	**ToCSV-[ZA:Mks22:07]**
African cassava mosaic Burkina Faso virus	ACMBFV	BF-Oua-127-08	HE616777	78,6	79,0
African cassava mosaic virus	ACMV	Cameroon/1998	J02057	75,5	75,6
Cassava mosaic Madagascar virus	CMMGV	Madgascar/Toliary/2006	HE617299	79,9	80,2
Chayote yellow mosaic virus	ChaYMV	Nigeria/Ibadan	AJ223191	77,5	77,3
	CLCuGeV/SD	Sudan/Gezira/1996/Sudan	AF260241	75,0	75,4

Cotton leaf curl Gezira virus	CLCuGeV/BF	Burkina Faso/Bazega/Okra/2009/Burkina Faso	FN554540	75,1	75,6
	CLCuGeV/Ca	Egypt/Cairo/Hollyhock/Cairo	AJ542539	74,6	75,5
	CLCuGeV/EG	Egypt/Aswan/Okra/Egypt	AF155064	75,3	75,6
	CLCuGeV/ML	Mali/Bamako/Okra/2006/Mali	EU024120	75,0	75,7
	CLCuGeV/NE	Niger/Niamey2/Okra/2007/Niger	EU432373	75,2	75,7
	CLCuGeV/Ok	Sudan/Okra/2007/Okra	FJ868828	74,9	75,2
Datura leaf curl virus	DaLCV	SD- Kha435-16	MF402918	75,0	75,5
Deinbollia mosaic virus	DMV	Tanzania-DB_T1A-2015	KT878829	75,0	75,9
East African cassava mosaic Kenya virus	EACMKV	Kenya/Mitaboni/K298/2002	AJ717572	76,1	76,4
East African cassava mosaic Malawi virus	EACMMV	Malawi/K/1996	AJ006460	76,1	76,6
East African cassava mosaic virus	EACMV/UG	Uganda/Severe 2/1997/Uganda	AF126806	75,5	75,6
	EACMV/KE	Kenya/Boa/K48/2001/Kenya	AJ717542	75,7	75,9
East African cassava mosaic Zanzibar virus	EACMZV	Tanzania/Uguja/1998	AF422174	78,2	77,9
Ocimum golden mosaic virus	OcGMV	UG-UG31-2015	MN313661	76,9	77,5
Okra yellow crinkle virus	OYCrV/ML	Mali/Bamako 2/2005/Mali	DQ875879	76,9	77,1
	OYCrV	Cameroon/Njombe 5/2007	HE793424	77,3	77,7
Pepper yellow vein Mali virus	PepYVMLV	Mali	AY502935	82,3	82,1
South African cassava mosaic virus	SACMV	South Africa	AF155806	78,6	79,3
Tobacco leaf curl Comoros virus	TbLCKMV	Comoros/Simboussa/2004	AM701760	78,1	79,0
	TbLCKMV	Comoros/Grande Comore, Foubouni	AM701762	78,2	79,2
Tobacco leaf curl Zimbabwe virus	TbLCZV	Zimbabwe	AF350330	83,8	84,5
	TbLCZV	Comoros/Grande Comore, Foumboudziouni	AM701756	83,0	83,6
Tomato curly stunt virus	ToCSV	South Africa/Onderberg/1998	AF261885	99,0	95,5
Tomato leaf curl Anjouan virus	ToLCAnV	Comoros/Ouani/2004	AM701758	78,0	77,9
Tomato leaf curl Arusha virus	ToLCArV	Tanzania/Kilimandjaro/2005	EF194760	79,7	80,4

Tomato leaf curl Burkina Faso virus	ToLCBFV	Burkina Faso-Loumbila-Tomate51B1-2013	KX853168	79,5	79,5
Tomato leaf curl Comoros virus	ToLCKMV	Mayotte/Kahani/2003	AJ865340	79,8	80,9
	ToLCKMV	Mayote/Dembeni/2003	AJ865341	79,6	80,6
Tomato leaf curl Diana virus	ToLCDiV	Madagascar/Namakely/2001	AM701765	79,3	79,6
Tomato leaf curl Ghana virus	ToLCGV	Ghana/Akumadan/2006	EU350585	78,5	78,3
	ToLCGV/Ag	Cameroon/AGFG24/2009/Ageratum	FR873228	78,4	78,3
	ToLCGV	GH-Aku-12	KT382326	80,1	80,0
	ToLCGV	Cote d'Ivoire/Kossou/TomatoCI104/2017	MN969998	78,8	78,5
	ToLCGV	Burkina Faso/Sossogona/Tomato82PA/2014	MH794632	79,0	79,3
	ToLCGV	Burkina Faso/Toussiana/Tomato479E7g/2016	MH794633	78,3	79,0
Tomato leaf curl Kunene virus	ToLCKunV	NA-2019	MT045996	78,9	79,0
Tomato leaf curl Madagascar virus	ToLCMGV/Men	Madagascar/Morondova/2001/Menabe	AJ865338	79,9	80,7
	ToLCMGV/Ats	Madagascar/Toliary/2001/Atsimo	AJ865339	79,1	79,9
Tomato leaf curl Mahe virus	ToLCMahV	SC-Pra-SC8-1b-17	MH410152	77,0	77,3
Tomato leaf curl Mali virus	ToLCMLV	Mali	AY502936	78,8	79,0
Tomato leaf curl Namakely virus	ToLCNaV	Madagascar/Namakely/2001	AM701764	79,9	80,4
Tomato leaf curl Nigeria virus	ToLCNGV	Nigeria/2006	FJ685621	79,8	79,8
Tomato leaf curl Seychelles virus	ToLCSCV	Seychelles/Val d'Endor/2004	AM491778	78,5	78,9
Tomato leaf curl Sudan virus	ToLCSDV/Gez	Sudan/Gezira/1996/Gezira	AY044137	76,8	76,9
	ToLCSDV	Sudan/WM/2011	KC763630	77,3	78,0
	ToLCSDV/Sha	Sudan/Shambat/1996/Shambat	AY044139	77,2	77,6
	ToLCSDV/YE	Yemen/2005/Yemen	EF110891	77,7	78,2
Tomato leaf curl Tanzania virus	ToLCTZV	TZ-Ten-05	DQ519575	78,8	79,6

Tomato leaf curl Toliara virus	ToLCToV	Madagascar/Miandrivazo/2001	AM701768	79,2	79,6
Tomato leaf curl Uganda virus	ToLCUV	Uganda/Iganga/2005	DQ127170	79,3	80,3
	ToLCUV	Kenya/Kiambu,Darugo	LC437667	80,4	81,4
Tomato yellow leaf curl Mali virus	TYLCMLV	Burkina Faso/Tom141/2013	LM651400	79,0	79,3
	TYLCMLV/ET	Ethiopia/Melkassa/2005/Ethiopia	DQ358913	79,2	79,9
	TYLCMLV/ML	Mali/2003/Mali	AY502934	79,0	79,2
West African Asystasia virus 2	WAAV2	West Africa/Asystasia2/2011	JF694486	76,7	76,4
Tomato yellow leaf curl Sardinia virus	TYLCSV	Italy/Sardinia/1988	X61153	75,5	75,3

S2 Table 2.4 Virus region-specific RT-PCR product sizes and PCR parameters used in this study.

RT-PCR product	**PCR annealing temperature (°C)**	**PCR extension time (s)**	**Product size (bp)**
V30 IR-vs	60	15	137
V30 IR-cs	60	15	137
V30 V2-vs	55	20	194
V22 V2-vs	55	20	129
V30 C6-cs	56	15	350
RNA 1-vs	56.1	20	494
RNA 1-cs	56.1	20	494
RNA 2-vs	58.1	20	504
RNA 2-cs	58.1	20	504
RNA 3-vs	60.1	60	1243
RNA 3-cs	60.1	60	1243

S2 Table 2.5. Predicted ORFs for ToCSV V30 and V22 DNA-A following standard code encoding proteins ≥40 aa in length.

ORF	Strand	Start codon nt position (V30/V22)	Stop codon nt position (V30/V22)	Protein length (aa) (V30/V22)
V1	+	299/298	1075/1074	258/258
V2	+	139(49^{N})/138	480/488	113(143^{N})/116
V3*****	+	1451/1450	1606/1605	51/51
V4*****	+	1612/1611	1740/1739	42/42
C1	-	2612(2648^{N})/2611(2647^{N})	1533/1532	359(371^{N})/359(371^{N})
C2	-	1624/1623	1217/1216	135/135
C3	-	1476/1475	1072/1071	134/134
C4	-	2455/2454	2198/2197	85/85
C5*****	-	610/609	428/427	60/60
C6*****	-	585/584	154/258	143/108
C7*****	-	515/514	378/377	45/45

Key: ^{N}, including putative 5' ORF extension; *****, putative ORF.

S2 Table 2.6 Predicted viral protein changes induced in V30ΔV2-s and V22ΔV2-s mutants following standard code.

Protein	Changes in V30ΔV2-s	Changes in V22ΔV2-s
CP (V1)	Ala36→Val (V30→V22 identity) Lys55→Arg (V30→V22 identity)	Val36→Ala (V22→V30 identity) Arg55→Lys (V22→V30 identity)
V2	Entire canonical V2 protein swapped to V22 V2, fused to putative V30 V2 N-terminal segment	Entire canonical V2 protein swapped to V30 V2, lacking putative V30 V2 N-terminal segment
C5*	Ile45→Val (V30→V22 identity) Leu50→Pro (V30→V22 identity)	Val45→Ile (V22→V30 identity) Pro50→Leu (V22→V30 identity)
C6*	Partial C6 swapped, resulting in protein identical to V22 C6.	Partial C6 swapped, resulting in protein identical to V30 C6.
C7*	Tyr13→Cys (V30→V22 identity)	Cys13→Tyr (V22→V30 identity)

Key: *, putative protein

Supplementary 3 (S3) Tables

S3 Table 3.1 Primers used in this study

Primer	Sequence (5'-3')	Purpose
ΔC1d(14)_FP	CTGTACTCAAGTCCCAATtAAGGTCCAGC ATCGAATA	Mutagenesis to generate V30ΔC1-d(14)
ΔC1d(14)_RP	TATTCGATGCTGGACCTTaATTGGGACTT GAGTACAG	
ΔC1d(22)_FP	CAGTGGCTaGTTGAGGGTGATGAAGAC	Mutagenesis to generate V30ΔC1-d(22)
ΔC1d(22)_RP	CACCCTCAACtAGCCACTGTACTCAAG	
ΔRepCd(14)_FP	CTCGCCGGACCTTaGTTCGGGCTGCTA	Mutagenesis to generate pCold I-RepC_14 and pMAL-c2X-RepC_14
ΔRepCd(14)_RP	TAGCAGCCCGAACtAAGGTCCGGCGAG	
V30M_FP	GAAACGACCAGTCTGAGGCT	PCR, sequencing (C1)
V30M2_FP	CGGTGCGTAAATCCATGGTT	PCR, sequencing (C1)
V30M_RP	TAAAGGCGGCATTCCCACTA	PCR, sequencing (C1)
pCold_FP	CCGACCATTTTCCTGTGCAA	PCR, sequencing (RepC)
pCold_RP	TTCACCGTCATCACCGAAAC	PCR, sequencing (RepC)
pMAL_FP	CCGACCATTTTCCTGTGCAA	PCR, sequencing (RepC)
pMAL_RP	GTTGTAAAACGACGGCCAGT	PCR, sequencing (RepC)
GAPDH_FP	CTGTTATTGGAGGAGGGAACAA	qPCR
GAPDH_RP	AGTCTTTCCTACCATGCCAAC	qPCR
V2_FP	CGACAGCCCGTTCACCAG	qPCR
V2_RP	GCCTGTACGTCCATGATCGTC	qPCR
V30C1N_FP	TGGTGGGTGTTTCAAGGTTT	cDNA synthesis, PCR (for RT-PCR)
V30C1N_RP	GCAATTTGAGAGTCAATCGGTA	PCR (for RT-PCR)

Key: lower case, target nt mutation

S3 Table 3.2 African mono- and bipartite begomovirus isolates used in this study.

Begomovirus species	Abbreviation	Isolate	DNA-A accession number
African cassava mosaic virus	ACMV	Cameroon/1998	J02057
Cotton leaf curl Gezira virus	CLCuGeV/ML	Mali/Bamako/Okra/2006/Mali	EU024120
Datura leaf curl virus	DaLCV	SD- Kha435-16	MF402918
East African cassava mosaic virus	EACMV/UG	Uganda/Severe 2/1997/Uganda	AF126806
Okra yellow crinkle virus	OYCrV/ML	Mali/Bamako 2/2005/Mali	DQ875879
Pepper yellow vein Mali virus	PepYVMLV	Mali	AY502935
South African cassava mosaic virus	SACMV	South Africa	AF155806
Tobacco leaf curl Comoros virus	TbLCKMV	Comoros/Simboussa/2004	AM701760
Tobacco leaf curl Zimbabwe virus	TbLCZV	Zimbabwe	AF350330
Tomato curly stunt virus	ToCSV	South Africa/Onderberg/1998	AF261885
Tomato leaf curl Anjouan virus	ToLCAnV	Comoros/Ouani/2004	AM701758
Tomato leaf curl Comoros virus	ToLCKMV	Mayotte/Kahani/2003	AJ865340
Tomato leaf curl Diana virus	ToLCDiV	Madagascar/Namakely/2001	AM701765
Tomato leaf curl Ghana virus	ToLCGV	Cote d'Ivoire/Kossou/TomatoCI104/2017	MN969998
Tomato leaf curl Kunene virus	ToLCKunV	NA-2019	MT045996
Tomato leaf curl Madagascar virus	ToLCMGV/Men	Madagascar/Morondova/2001/Menabe	AJ865338
Tomato leaf curl Mahe virus	ToLCMahV	SC-Pra-SC8-1b-17	MH410152
Tomato leaf curl Mali virus	ToLCMLV	Mali	AY502936
Tomato leaf curl Namakely virus	ToLCNaV	Madagascar/Namakely/2001	AM701764
Tomato leaf curl Nigeria virus	ToLCNGV	Nigeria/2006	FJ685621
Tomato leaf curl Seychelles virus	ToLCSCV	Seychelles/Val d'Endor/2004	AM491778
Tomato leaf curl Sudan virus	ToLCSDV/Sha	Sudan/Shambat/1996/Shambat	AY044139
Tomato leaf curl Toliara virus	ToLCToV	Madagascar/Miandrivazo/2001	AM701768
Tomato leaf curl Uganda virus	ToLCUV	Uganda/Iganga/2005	DQ127170
Tomato yellow leaf curl Mali virus	TYLCMLV	Burkina Faso/Tom141/2013	LM651400
Tomato yellow leaf curl Sardinia virus	TYLCSV	Italy/Sardinia/1988	X61153