Characterization of the banana streak virus capsid protein and mapping of the immunodominant continuous B-cell epitopes to the surface-exposed N terminus

Jenny N. Vo,[1,2,3]† Paul R. Campbell,[1,4] Nur N. Mahfuz,[3] Ras Ramli,[3] Daniel Pagendam,[5] Ross Barnard[3] and Andrew D. W. Geering[1,2]

Correspondence
Jenny N. Vo
jennynhu.vo@gmail.com

[1]Plant Biosecurity Cooperative Research Centre, LPO Box 5012, Bruce, Australian Capital Territory 2617, Australia

[2]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, GPO Box 267, Brisbane, Queensland 4001, Australia

[3]School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, Queensland 4072, Australia

[4]Queensland Department of Agriculture, Fisheries and Forestry, GPO Box 267, Brisbane, Queensland 4001, Australia

[5]CSIRO Mathematics, Informatics and Statistics, Ecosciences Precinct, 41 Boggo Road, Dutton Park, Queensland 4102, Australia

This study identified the structural proteins of two badnavirus species, *Banana streak MY virus* (BSMYV) and *Banana streak OL virus* (BSOLV), and mapped the distribution of continuous B-cell epitopes. Two different capsid protein (CP) isoforms of about 44 and 40 kDa (CP1 and CP2) and the virion-associated protein (VAP) were consistently associated with purified virions. For both viral species, the N terminus of CP2 was successfully sequenced by Edman degradation but that of CP1 was chemically blocked. *De novo* peptide sequencing of tryptic digests suggested that CP1 and CP2 derive from the same region of the P3 polyprotein but differ in the length of either the N or the C terminus. A three-dimensional model of the BSMYV-CP was constructed, which showed that the CP is a multi-domain structure, containing homologues of the retroviral capsid and nucleocapsid proteins, as well as a third, intrinsically disordered protein region at the N terminus, henceforth called the NID domain. Using the Pepscan approach, the immunodominant continuous epitopes were mapped to the NID domain for five different species of *banana streak virus*. Anti-peptide antibodies raised against these epitopes in BSMYV were successfully used for detection of native virions and denatured CPs in serological assays. Immunoelectron microscopy analysis of the virion surface using the anti-peptide antibodies confirmed that the NID domain is exposed on the surface of virions, and that the difference in mass of the two CP isoforms is due to variation in length of the NID domain.

INTRODUCTION

The *Caulimoviridae* is the only family of viruses in the plant kingdom with a dsDNA genome and has a recent common ancestor with the *Metaviridae* and *Retroviridae* (King *et al.*, 2012; Xiong & Eickbush, 1990). These viruses possess a conserved core replicon comprising genes for one or more capsid proteins (CPs) (alternatively called 'group-specific antigen', or *gag*, in retroviruses), an aspartic protease (AP) enzyme and a reverse transcriptase (RT) enzyme containing a ribonuclease H1 (RH1) domain (King *et al.*, 2012). The genus *Badnavirus* is overwhelmingly the largest of the eight recognized genera in the family *Caulimoviridae*, with nearly twice the number of species as the combined total of all other genera (Virus Taxonomy: 2015 Release, www.ictvonline.org). Badnaviruses are

particularly prominent in tropical regions; at least nine different badnavirus species are known to infect banana (*Musa* spp.), two of the most widely distributed being *Banana streak MY virus* (BSMYV) and *Banana streak OL virus* (BSOLV) (Geering *et al.*, 2005; Harper & Hull, 1998).

Badnaviruses have bacilliform-shaped virions, of 30×130–150 nm, and a circular, non-covalently closed, dsDNA genome of 7.2–9.2 kbp (King *et al.*, 2012). The badnavirus genome has three ORFs that are all present in the plus strand: ORF1 encodes a protein of unknown function (protein P1); ORF2, the virion-associated protein (VAP; protein P2); and ORF3, a P3 polyprotein that contains the precursors of the movement protein (MP), CP, AP and RT-RH1. By inference from better studied members of the *Caulimoviridae* and *Retroviridae*, it is thought that the AP autocleaves from the P3 polyprotein and then processes the other precursors to give rise to the mature proteins (Debouck *et al.*, 1987; Laco & Beachy, 1994). The AP substrate sites have not been determined for any badnavirus species, and these cannot be predicted through searches for conserved amino acid sequence motifs, as substrate shape is probably the most important factor affecting the specificity of recognition for the retroviral aspartic protease protein family (Marmey *et al.*, 2005; Prabu-Jeyabalan *et al.*, 2002; Tözsér, 2010).

The mature virion of *Commelina yellow mottle virus* (ComYMV), the type species of the genus *Badnavirus*, contains two types of protein, namely the CP and the VAP (Cheng *et al.*, 1996). Additionally, the immature virion is thought to contain the P1 protein, although it is undetectable after chloroform treatment, perhaps because of selective disruption of the immature virions (Cheng *et al.*, 1996). For *Cauliflower mosaic virus* (CaMV), the type species of the family *Caulimoviridae*, the VAP is non-essential for virion morphogenesis (Kobayashi *et al.*, 2002) and is non-dectectable in purified virions but, when added, decorates the virion surface in a stoichiometric ratio of 1 : 1 with the capsomers. This VAP has two antiparallel coiled-coil domains at its N terminus and self-associates to form dimers, which then assemble to form a triskelion structure cementing three hexavalent or pentavalent capsomers together (Hoh *et al.*, 2010; Plisson *et al.*, 2005). The C terminus of this VAP has a nucleic acid binding domain (Jacquot *et al.*, 1996, 1997; Mougeot *et al.*, 1993) and penetrates through pores between the capsomers and is thought to associate with the genomic DNA. The CaMV VAP is multifunctional, as it also interacts with the P2 protein to form a molecular bridge between the virion and the stylet of its aphid vector (Martinière *et al.*, 2013), and facilitates cell-to-cell movement (Stavolone *et al.*, 2005).

Among the *Caulimoviridae*, a general observation is the presence of two or three differently sized CP isoforms in purified virion preparations (Cheng *et al.*, 1996; Leclerc *et al.*, 1999; Marmey *et al.*, 1999; Qu *et al.*, 1991; Vo *et al.*, 2015). Marmey *et al.* (1999) concluded that the smaller of the two CP isoforms of *Rice tungro bacilliform virus* (RTBV)

was a degradation product of the larger isoform, as it was only present in purified virion preparations and not in infected leaf extracts, although this result was contradicted by Hay *et al.* (1994). A possible explanation for this contradiction was provided by Cheng *et al.* (1996), who found that the smaller CP of ComYMV, which is closely related to RTBV, was undetectable in newly emerged, systemically infected leaves but became more abundant as the leaf aged. For CaMV, targeted degradation of the acidic N terminus does occur during virion maturation, and this cleavage allows the virion to be trafficked to the nuclear pore where disassembly occurs and genome replication begins (Champagne *et al.*, 2004; Karsies *et al.*, 2001).

Other than the aforementioned work on ComYMV, the structural proteins of badnaviruses are poorly characterized, probably due to the technical difficulties associated with working with these viruses. There are also worldwide shortages or absences of antisera to nearly every badnavirus species. In this study, the structural proteins of BSMYV and BSOLV have been characterized using a combination of proteomics and bioinformatics approaches. The locations of continuous B-cell epitopes on these proteins have then been identified by functional epitope mapping and chemically synthesized peptides mimicking these epitopes have been demonstrated to be strong immunogens. Finally, empirical support for a three-dimensional (3D) CP model has been provided by probing the virion with antibodies made to peptides containing the epitopes.

RESULTS

Identification of the capsid and virion-associated proteins

Using RaptorX, the region of the BSMYV P3 polyprotein spanning amino acid residues E_{499}–E_{698} was predicted (E-value=2.3×10^{-9}) to belong to the retrovirus Gag P30 core shell protein family (Pfam: PF02093). The same domain was identified in the BSOLV P3 polyprotein, spanning E_{472}–E_{671}. In this and a previous study, protein doublets in the size range of 40–44 kDa were detected when purified preparations of BSMYV and BSOLV were analysed by SDS-PAGE, although the quantities of the larger protein did vary between experiments and was sometimes difficult to detect (Fig. 1; Vo *et al.*, 2015). These proteins were hypothesized to be different isoforms of the CP, hereafter referred to as CP1 (large) and CP2 (small). To identify these proteins, peptide fragments obtained by trypsin digestion were sequenced by tandem MS and all peptide fragments were matched six to nine sequence blocks that were N-terminal of the zinc finger motif in the P3 polyprotein (Fig. 2). Furthermore, the array of sequenced peptides for CP1 was virtually identical to that of CP2, indicating that they derive from the same part of the P3 polyprotein but differ in mass due to differences in length of either the N or the C terminus. An additional protein of about 15–18 kDa was clearly visible in fresh virion preparations (Fig. 1),
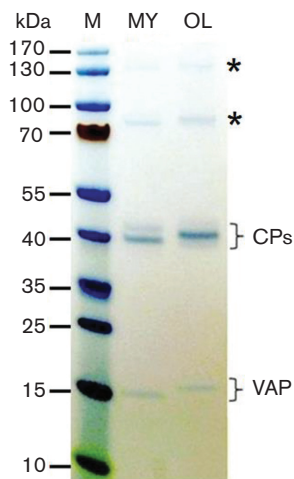
Fig. 1. Detection of the structural proteins of BSMYV (MY) and BSOLV (OL) by PAGE under denaturing conditions. Protein bands of about 40–44 and 15 kDa are the major capsid proteins (CPs) and the virion-associated protein (VAP), respectively, as determined by *de novo* peptide sequencing. Asterisks denote possible polyprotein P3 intermediates containing the CP. Lane M is the PageRuler Prestained Protein Ladder (Fermentas).

which was identified to be the VAP using the same methods (Fig. 2).

For both BSV species, the CP1 had a blocked N terminus, preventing Edman degradation sequencing. However, the N terminus of the CP2 isoform was determined to be S_{481} TAPD and S_{426}YRPPD for BSMYV and BSOLV, respectively. Interestingly, the N terminus of the BSMYV CP2 mapped to almost the same position as that of RTBV relative to the N terminus of the Gag-p30 domain (Fig. S1, available in the online Supplementary Material).

3D modelling and structural analysis of the BSMYV CP

When using I-TASSER for 3D modelling, it first attempts to retrieve template proteins of similar folds (or super-secondary structures) from the Protein Data Bank (PDB) by LOMETS threading. During this initial threading stage, the putative nucleic acid binding domain of the BSMYV CP2 (K_{755}–E_{797}), which contains a zinc finger motif, aligned with the nucleocapsid (NC) protein of several retroviruses (i.e. pdb templates: 1a1t, 1aaf, 2ec7, 1a6b, 1u6p, 2ihx). A second domain (P_{504}–R_{716}) comprising ten consecutive α-helices, also aligned with the capsid (CA) protein of several retroviruses (i.e. pdb: 3gv2, 3h47, 2eia, 3tir and 1l6n) and the myosin V cargo binding domain of *Saccharomyces cerevisiae* (pdb: 2f6h). Apart from the retroviral gag homologues, structural templates of other non-structural proteins (e.g. catalytic or regulatory proteins) such as a region of intersectin (pdb: 1ki1), sulfotransferase (pdb: 4gbm), DNA polymerase (pdb: 1ih7) and acid phosphatase (pdb: 1qfx) were among the top ten templates integrated

by I-TASSER to build the 3D structural prototypes of the entire CP2. These domain matching templates had normalized Z-scores of >1.1, suggesting good-quality alignments (Roy *et al.*, 2010).

Five models were generated for the CP2 and all had *C*-scores within the normal range of −5 to 2 (Roy *et al.*, 2010). A *C*-score greater than −1.5 indicates the probability that the model prediction is correct is greater than 90 % (Roy *et al.*, 2010) but no model met this stringent criterion. However, when the structural predictions were limited to either the CA or the NC domain, scores of −1.92 and −1.0, respectively, were obtained (data not shown).

Two intrinsically disordered protein regions (IDPRs) were identified in the CP2, which could explain some of the difficulties in modelling the entire protein (Fig. 3). The N-terminal intrinsically disordered (NID) domain contains a high frequency of alanine, proline and serine and although the former two amino acids possess hydrophobic side chains, the domain is overall hydrophilic in nature, suggesting that it protrudes from the surface of the virion. Two putative casein kinase type II (CKII) phosphorylation sites, LGS_{482}TAPD and KWK_{496}SPTE, and a protein kinase C phosphorylation site, KG_{491}SFK, were present in the NID domain. The second, lysine-rich (basic) IDPR was at the front of the zinc finger motif in the NC domain, and contained a putative nuclear localization signal (Fig. 3). The overall structural prediction and analysis showed that the BSV CP is a multi-domain structure containing the protruding NID domain, and homologues of the retroviral CA and NC proteins (Fig. 4).

The arrangement of secondary structural elements in the BSMYV CP is shared by other viruses in the *Caulimoviridae* such as BSOLV and RTBV (Fig. S2). The CaMV CA domain differs only through the presence of an extra α-helix at the C terminus. No acidic spacer sequence could be found between the CA and NC domains (Fig. 4), as occurs for members of the *Metaviridae* (Sandmeyer & Clemens, 2010).

Distribution of continuous epitopes

When the BSMYV CP peptide library was probed with the JVQ IgGs in the Pepscan assay, eight statistically significant ($P<0.05$) absorbance peaks were observed, each representing a continuous B-cell epitope (Fig. 5). The two largest peaks were in the NID domain, and corresponded to peptides P4–5 [E1] and P9–10 [E2]. Peptide 18 [E3] represented the apex of the third highest peak, and straddled the boundary between the NID and CA domains, with the five C-terminal residues (A_{512}MFVM) predicted to form part of a β-strand that is buried inside the CP (Fig. 5). The remaining absorbance peaks (peptides P25–P27 [E4], P40–P41 [E5], P53–P54 [E6], P65 [E7], P77–P79 [E8]), which mapped to either the CA or the NC domains, were relatively minor in scale compared to those in the NID domain and therefore were not considered in later investigations. Interestingly, epitope E1 mapped to a region of the protein that
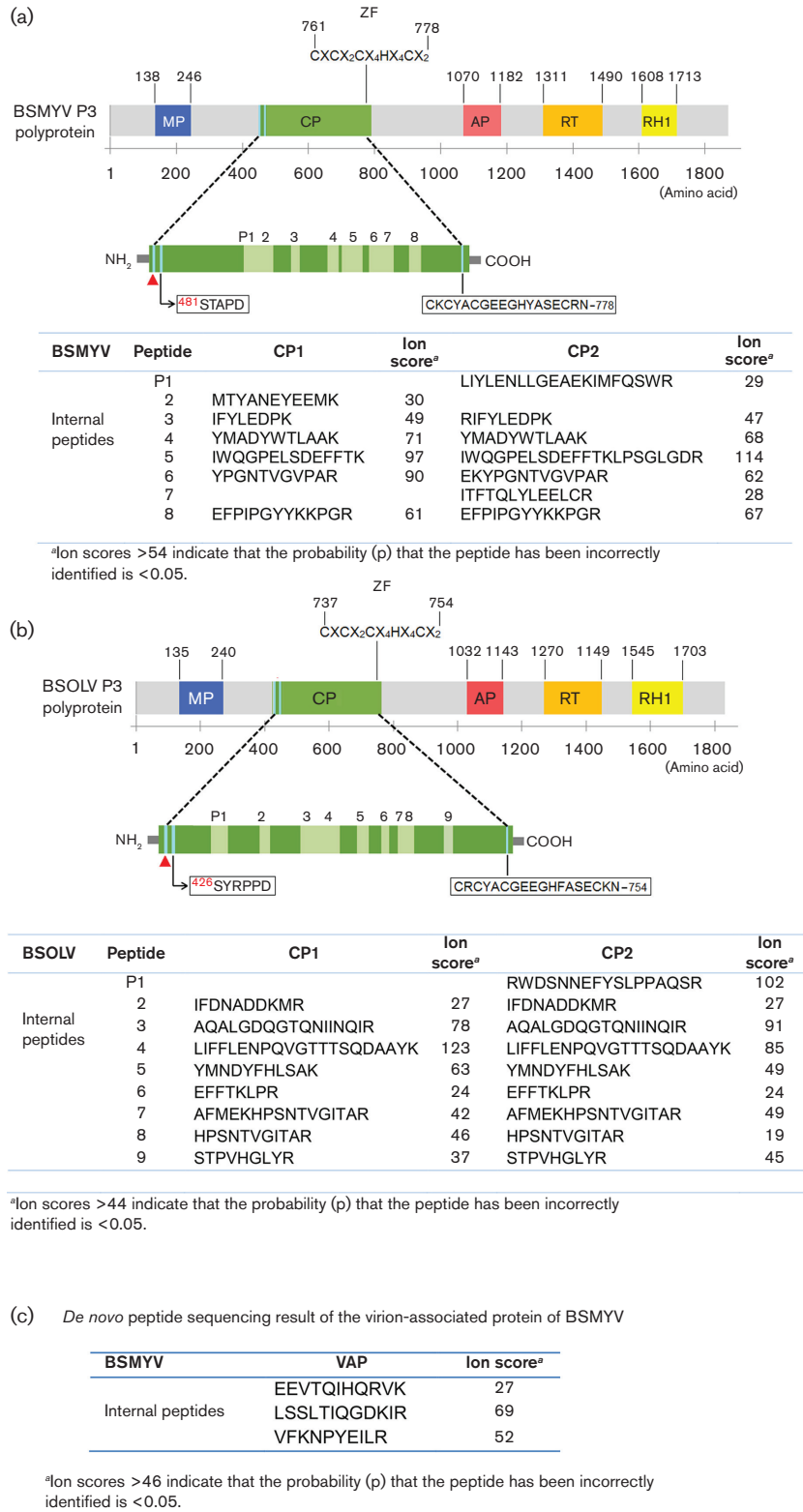
(a)

BSMYV P3 polyprotein — domains: MP (138–246), CP, ZF (761–778) CXCX₂CX₄HX₄CX₂, AP (1070–1182), RT (1311–1490), RH1 (1608–1713)

481 STAPD ... CKCYACGEEGHYASECRN-778

BSMYV	Peptide	CP1	Ion score[a]	CP2	Ion score[a]
	P1			LIYLENLLGEAEKIMFQSWR	29
	2	MTYANEYEEMK	30		
Internal peptides	3	IFYLEDPK	49	RIFYLEDPK	47
	4	YMADYWTLAAK	71	YMADYWTLAAK	68
	5	IWQGPELSDEFFTK	97	IWQGPELSDEFFTKLPSGLGDR	114
	6	YPGNTVGVPAR	90	EKYPGNTVGVPAR	62
	7			ITFTQLYLEELCR	28
	8	EFPIPGYYKKPGR	61	EFPIPGYYKKPGR	67

[a]Ion scores >54 indicate that the probability (p) that the peptide has been incorrectly identified is <0.05.

(b)

BSOLV P3 polyprotein — domains: MP (135–240), CP, ZF (737–754) CXCX₂CX₄HX₄CX₂, AP (1032–1143), RT (1270/1149–1545), RH1 (1703)

426 SYRPPD ... CRCYACGEEGHFASECKN-754

BSOLV	Peptide	CP1	Ion score[a]	CP2	Ion score[a]
	P1			RWDSNNEFYSLPPAQSR	102
	2	IFDNADDKMR	27	IFDNADDKMR	27
Internal peptides	3	AQALGDQGTQNIINQIR	78	AQALGDQGTQNIINQIR	91
	4	LIFFLENPQVGTTTSQDAAYK	123	LIFFLENPQVGTTTSQDAAYK	85
	5	YMNDYFHLSAK	63	YMNDYFHLSAK	49
	6	EFFTKLPR	24	EFFTKLPR	24
	7	AFMEKHPSNTVGITAR	42	AFMEKHPSNTVGITAR	49
	8	HPSNTVGITAR	46	HPSNTVGITAR	19
	9	STPVHGLYR	37	STPVHGLYR	45

[a]Ion scores >44 indicate that the probability (p) that the peptide has been incorrectly identified is <0.05.

(c) *De novo* peptide sequencing result of the virion-associated protein of BSMYV

BSMYV	VAP	Ion score[a]
	EEVTQIHQRVK	27
Internal peptides	LSSLTIQGDKIR	69
	VFKNPYEILR	52

[a]Ion scores >46 indicate that the probability (p) that the peptide has been incorrectly identified is <0.05.

Fig. 2. *De novo* peptide sequencing results for (a) the capsid protein (CP) of BSMYV, (b) the CP of BSOLV and (c) the virion-associated protein (VAP) of BSMYV. A red triangle denotes the estimated position of the blocked N terminus of the CP1 isomer. The position of a conserved CCHC-zinc finger (ZF) at the C terminus of the CP is indicated. Other domains within the P3 polyprotein are the movement protein (MP), aspartic protease (AP), reverse transcriptase (RT) and RNaseH1 (RH1).
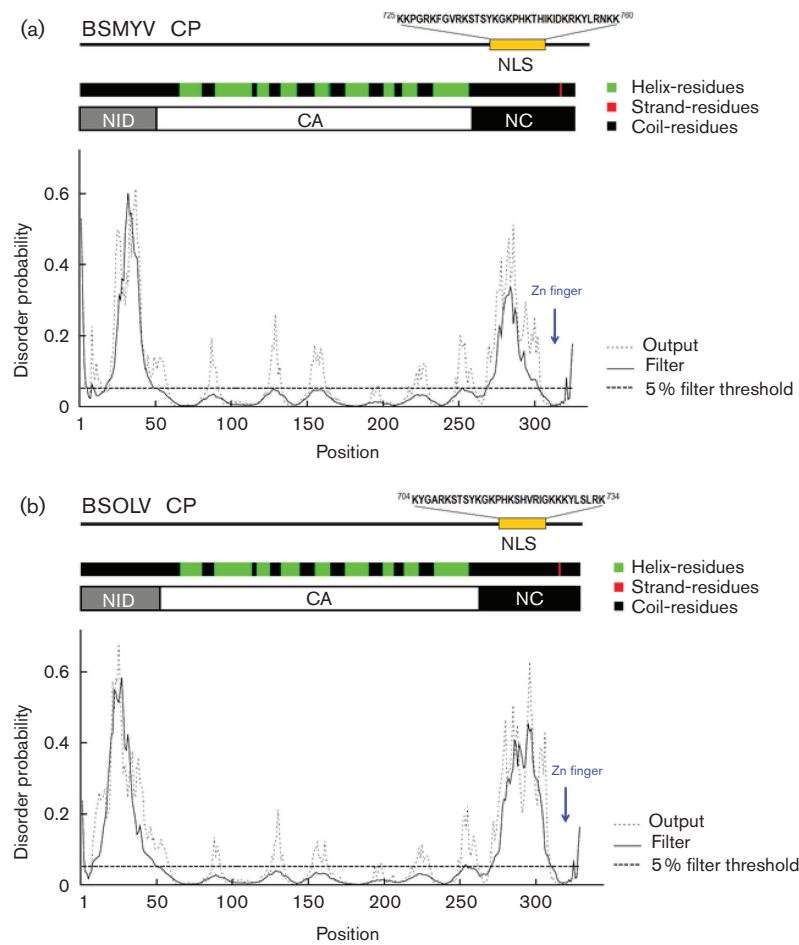
Fig. 3. Disorder profiles for the capsid proteins (CPs) of (a) BSMYV and (b) BSOLV. DISOPRED2 was used to predict protein disorder with a false positive rate threshold of 2 %. NID is the abbreviation for N-terminal intrinsically disordered domain; CA, the capsid domain and NC, the nucleocapsid domain. The positions and sequences of putative nuclear localization signals (NLS) in the NC domains, as predicted using NLStradamus (threshold set at 0.6), are also shown.
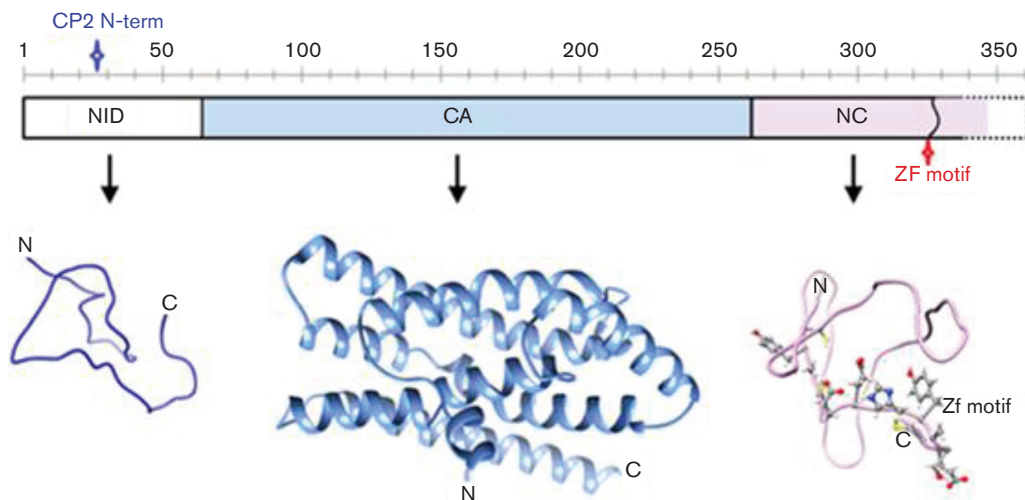

Fig. 4. Structural model for the BSMYV CP. CA, capsid domain; NC, nucleocapsid domain; ZF, zinc finger motif.
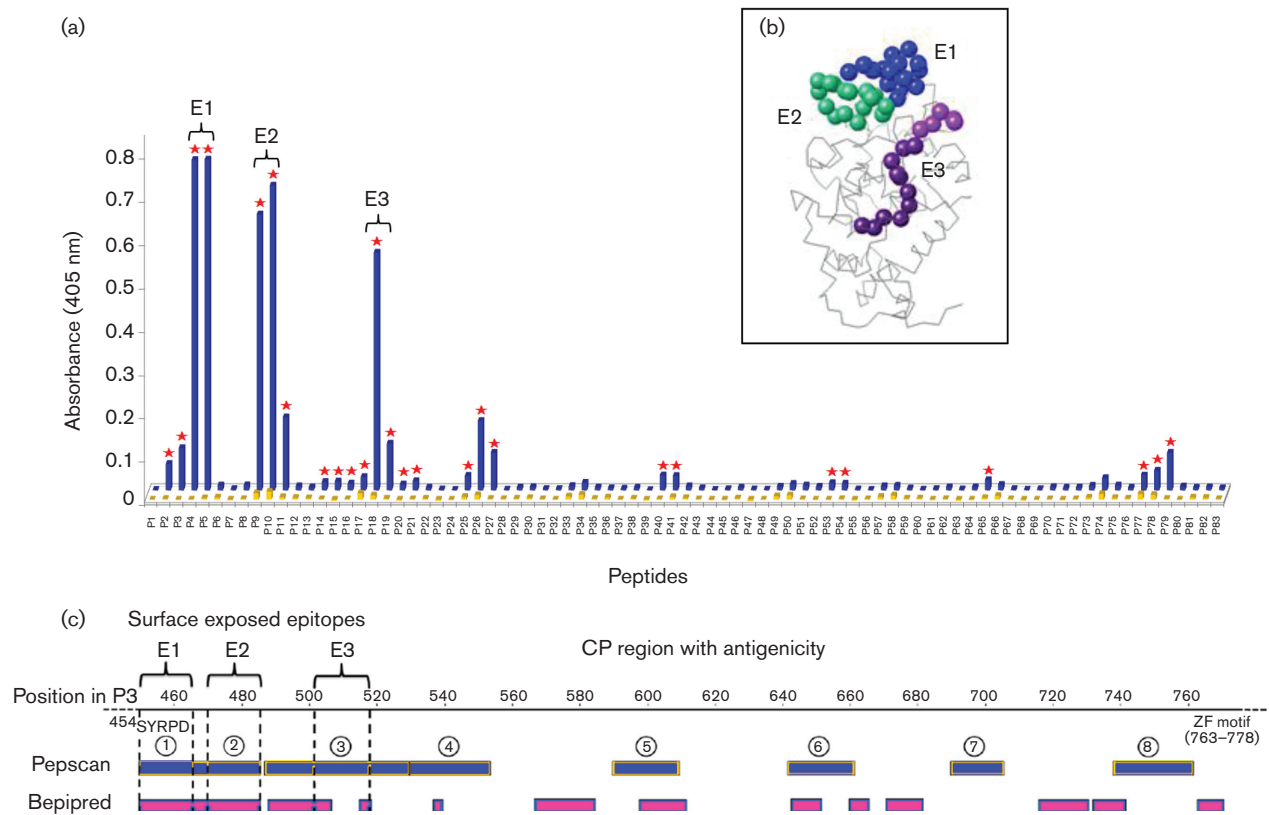
Fig. 5. Distribution of continuous epitopes on the BSMYV CP. (a) Pepscan profile using library of overlapping 16-mer peptides spanning the entire CP; peptides are listed in Table S1. Blue and yellow bars represent ELISA absorbance values for the seroconverted (JVQ1) and naïve IgG antibodies, respectively, and a red star indicates that the treatments are significantly different ($P<0.05$). (b) Positions of major epitopes (E1, E2 and E3) on the 3D model of the CP. (c) Continuous B-cell epitopes predicted using the Pepscan method (blue bars) and BepiPred method (pink bars) (threshold=0.35 and average score 0.25).

was upstream of the N terminus of the CP2 isoform, suggesting that the additional size of the CP1 isoform was due to an extension of the NID domain.

Pepscan profiles were also generated for the CPs of BSOLV, *Banana streak GF virus* (BSGFV), *Banana streak IM virus* (BSIMV) and *Banana streak CA virus* (BSCAV) but using a reduced repertoire of peptides covering only the NID domain and the N terminus of the CA domain. Homologous antisera were unavailable for these BSV species, but instead, the peptides were probed with IgGs from the PMxR2-2C antiserum, which was made against a mixture of uncharacterized badnavirus isolates from 32 sugarcane cultivars and banana cv. Mysore, and is known to cross-react with at least five serotypes of BSV (Ndowora & Lockhart, 2000). As with the BSMYV CP, highly reactive continuous epitopes were identified in the NID domains of these BSV species (Fig. S3).

The contribution of the VAP to the antigenic profile of BSMYV was also investigated using the JVQ1 IgGs. Antibodies that recognized most peptides were present in the seroconverted serum but peptide VAP$_{41-56}$

(LTRQLNTLIYSVVKIK) had a significantly higher absorbance value than the others (Fig. S4). However, the titre of cross-reacting antibodies was much lower than with the three N-terminal epitopes on the CP. VAP$_{41-56}$ mapped to the second of four α-helices of the predicted secondary structure of the VAP.

Finally, the results obtained by Pepscan analysis of the BSMYV CP were compared with those obtained using BepiPred 1.0, a program that predicts the location of continuous B-cell epitopes using a combination of a hidden Markov model and a propensity scale method. Overall, there was a good correlation between the *in silico* predictions and the empirical results, and most importantly, the highest scoring epitopes were in the NID domain (Fig. 5).

Immunogenicity of synthetic peptides mimicking the CP epitopes

Peptides P5, P10 and P18 from the NID domain of the BSMYV CP (Fig. 5, Table S1) were considered the best candidates to be immunogenic and capable of producing
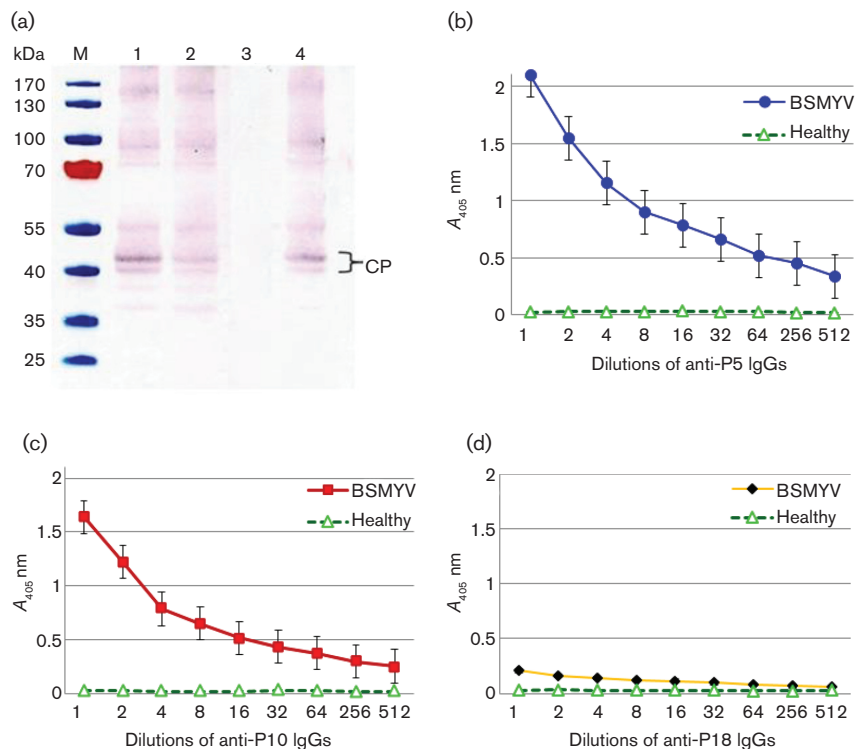
Fig. 6. Detection of BSMYV by Western blot assay and plate-trapped antigen (PTA) ELISA using anti-peptide immunoglobulin G (IgG) antibodies. (a) Western blot assay of partially purified BSMYV (virus miniprep). Lane M is the prestained protein marker and lanes 1, 2 and 3 correspond to the anti-P5, P10 and P18 antibodies, while lane 4 is the JVQ1 antibodies. (b)–(d) Titration curves for the anti-P5 (blue line), anti-P10 (red line) and anti-P18 (yellow line) antibodies in PTA ELISA, using a crude virus extract (miniprep) as the antigen. The starting concentration for the dilution series was 20 μg ml^{-1}. For the negative control, a leaf extract from healthy banana was used. Each data point represents the mean absorbance value at 405 nm (A_{405}) of three replicates, after 35 min of substrate incubation at 37 °C.

antibodies that recognize their cognate epitopes on the native antigen. The reactivity of the antibodies to their corresponding synthetic peptides was tested by plate-trapped antigen (PTA) ELISA by the manufacturer and each had a similar titre, >1 : 64 000 (data not shown). In PTA ELISA and Western blot assay using crude virion extracts (virus miniprep) as antigen, the anti-P5 and anti-P10 antibodies successfully recognized BSMYV, but the anti-P18 antibodies did not (Fig. 6). Negligible reactions were observed with miniprep of healthy banana tissue.

As expected, both CP isoforms were detected by Western blot assay using the anti-P10 antibodies, reflecting the fact that nine of the amino acid residues within the P10 peptide correspond to the N terminus of the CP2. An unexpected result was that the anti-P5 antibodies crossreacted with the CP2 in Western blot assay, as the P5 peptide sequence is hypothesized to be present only in the extended NID domain of CP1. However, when peptides 5 and 10 were chemically conjugated to the carrier protein BSA, and analysed by Western blot assay, the anti-P5 antibodies cross-reacted with both the P5 and the P10 peptide–BSA conjugates but not with pure BSA,

whereas the anti-P10 antibodies only reacted with its homologous peptide–BSA conjugate (Fig. S5). This crossreactivity of anti-P5 antibody to the P10 peptide in Western blot was not evident in double antibody sandwich (DAS) ELISA, suggesting that the Western blot results were due to a conformational effect from the denaturing conditions of SDS-PAGE.

A small number of proteins that were both larger and smaller in mass than the CPs were detected by Western blot assay using the anti-peptide antibodies and a virtually identical protein profile was detected with the JVQ1 antibodies that were prepared against whole virions. It is hypothesized that the larger proteins are P3 polyprotein processing intermediates (Cheng *et al.*, 1996; Vo *et al.*, 2015) and the smaller proteins may be degradation products.

To further examine the hypothesis that the NID domain is surface-exposed, antibody decoration tests were done. Strong decoration of the virion was observed using both the anti-P5 and the anti-P10 antibodies at intensities equivalent to that observed with the JVQ1 antibodies (Fig. 7). Since the CP N-terminal domain is the most variable region
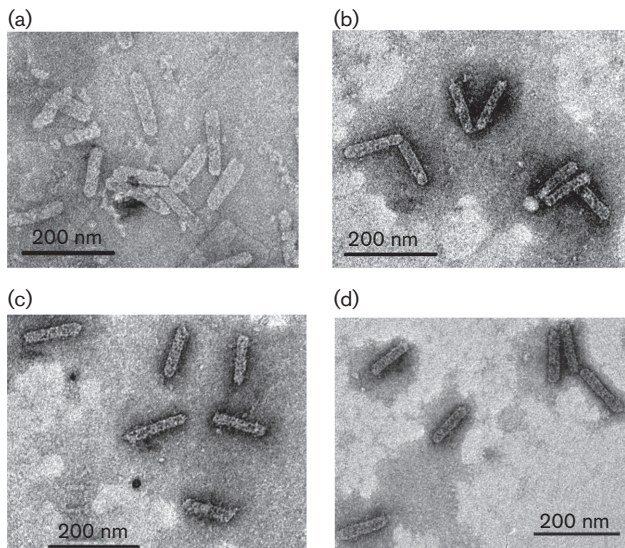
Fig. 7. Decoration of BSMYV with anti-peptide antibodies. Electron micrographs showing (a) undecorated virions, (b) JVQ1 antibodies (positive control), (c) anti-P5 antibodies and (d) anti-P10 antibodies.

among BSVs based on protein sequence analysis, it was expected that functional epitopes identified at the extreme N-terminal sequence of the BSV CP give rise to species-specific anti-peptide antibodies. The anti-peptide antibodies only detected the BSMYV CPs but not those from BSOLV, BSIMV or BSGFV in a Western blot assay (data not shown).

DISCUSSION

In this study, we have characterized the structural proteins of BSMYV and BSOLV and mapped the location of continuous epitopes on the CP of these and three other BSV species. The CP core structural domains were modelled with a high level of confidence, but the complete CP with relatively less confidence, due to the presence of IDPRs in the NID and NC domains. IDPRs cannot be adequately described by a single equilibrium 3D structure (Tompa & Fuxreiter, 2008; Uversky, 2013). Rather, according to the 'induced-fit' hypothesis, the structure of IDPRs is malleable and dependent on the shape of the interacting protein molecule or ligand (Dyson & Wright, 2005; Uversky, 2013). IDPRs have structural heterogeneity and this conformational plasticity endows them with a greater potential for multi-functionality. When compared to ordered protein regions, IDPRs interact more rapidly with partner molecules, are efficiently regulated by targeted degradation or proteolysis, have a high accessibility of sites for post-translational modification, and have the ability to bind to several structurally diverse partners, with which they often form specific but weak

complexes, favouring a regulatory function (Uversky, 2013).

Two major CP isoforms were detected, which derive from the same region of the P3 polyprotein but differ in length of either the N or the C terminus (Cheng *et al.*, 1996; Leclerc *et al.*, 1999; Qu *et al.*, 1991). The N terminus of the larger CP isoform could not be sequenced as the protein appeared to be chemically blocked, a problem previously experienced with CaMV, RTBV and *Soybean chlorotic mottle virus* (Marmey *et al.*, 1999; Martinez-Izquierdo & Hohn, 1987; Takemoto & Hibi, 2001). However, from the epitope mapping experiments, together with antibody decoration tests done using anti-peptide antibodies, it can be concluded that the differences in mass are due to variation in length of the NID domain. We hypothesize that the two CP isoforms are naturally occurring as a consequence of the presence of alternative aspartic protease cleavage sites or because of targeted degradation of the N terminus rather than because of proteolysis during purification. The two isoforms may play different roles during the replication cycle, such as regulating the partitioning of the CP between sites of virion assembly in the cytoplasm and disassembly at the nuclear pore (Karsies *et al.*, 2001, 2002). Two putative CKII phosphorylation sites were identified in the extended 27 aa N terminus of the CP1 isoform. CaMV has three serine residues in the acidic terminus of the CP precursor that are phosphorylated by a CKII enzyme, and mutation of all three residues abolishes infectivity (Chapdelaine *et al.*, 2002). A favoured hypothesis is that phosphorylation regulates the RNA or DNA binding activity of the CP precursor, which allows the specific packaging of the RNA pregenome to form the reverse transcription complex (Guerra-Peraza *et al.*, 2000).

A second basic IDPR was identified in the N-terminal side of the zinc finger motif of the BSV CP and contains a putative nuclear localization signal (Fig. 3). The homologous region of the RTBV CP contains a functional nuclear localization signal (NLS) and an importin α binding motif (Martinez-Izquierdo & Hohn, 1987), suggesting similarities between RTBV and badnaviruses in the mechanism of transport of the virus genome into the nucleus. This basic IDPR is also likely to be a determinant of the specificity of interaction between the CP and a putative encapsidation signal at the top of the hairpin in the pregenomic RNA leader (Geering *et al.*, 2005; Guerra-Peraza *et al.*, 2000).

Using the Pepscan method (Geysen *et al.*, 1984), highly reactive epitopes were identified in the NID domain of the five BSV species that were examined. The fact that this domain is a major antigenic determinant for the viruses is concordant with theory on epitope structure, as 85 % of well-characterized epitopes that are described in the scientific literature have a linear stretch of five or more residues and the epitopes are mainly found in surface-exposed or protruding regions of the protein (Kringelum *et al.*, 2013). Of the three reactive N-terminal epitopes that were identified, two were successfully used to produce anti-peptide antibodies that were capable of

binding their cognate peptide fragments in the CP in Western blot and antibody decoration assays. Furthermore, these antibodies were efficiently used as detecting antibodies in double sandwich ELISA (data not shown). The anti-peptide antibodies to the third epitope, E3, were poorly reactive with the CP in these same assays. In the 3D structural model of the BSMYV CP, epitope E3 was partially buried in the protein. The identification of this epitope could be explained by the anti-BSMYV polyclonal antiserum containing a population of IgGs that have been produced in response to parts of the CP being exposed by processing of the antigen in the immunized rabbit. Alternatively, the poor immunogenicity of the linear peptide may be due to the fact that it only partially mimics the epitope present in the CP.

Empirical methods for epitope identification are both expensive and not particularly feasible for viruses that occur in remote regions, such as some of the BSV species that are restricted to central Africa (Harper et al., 2005). Using a combination of bioinformatics methods such as DISOPRED2 to identify the NID domain on the CP, and BepiPred to predict continuous B-cell epitopes, it would be possible to identify immunogenic peptides to make antisera, while only having the conceptual amino acid sequence at hand. Ultimately, peptides mimicking viral epitopes could provide an unlimited source of homologous immunogens for the production of peptide-based immunodiagnostic reagents, particularly when the antigen is difficult to obtain or purify.

In summary, we have defined and characterized the structural proteins of BSMYV and BSOLV for the first time. These results will have an applied outcome by providing information that could be used to more rapidly and efficiently develop immunodiagnostic reagents, as well as providing a structural model of the CP that will help better understand how the virus regulates its replication and interacts with its cellular environment and insect vectors. The identified epitopes were useful for verifying the computationally predicted structure of BSMYV CP. Finally, our investigation presages future applications of the anti-peptide antibodies for the development of a range of serological assays for badnaviruses in general.

METHODS

Virus isolates and purification methods. BSMYV and BSOLV were maintained in *Musa* spp. cv. Mysore (AAB Group) and TMB× hybrid (AAB group) plants, respectively, grown in an irrigated field plot in Brisbane, Australia.

Virus purifications and minipreps (small-scale, crude virus extracts) were done as described previously (Vo et al., 2015). As a proxy for virion concentration, total proteins were quantified using a Quant-iT Protein Assay Kit and a Qubit Fluorometer (Life Technologies) according to the manufacturer's protocol.

Antisera and immunoglobulins. The JVQ1 rabbit polyclonal antiserum, prepared against purified BSMYV, was as described by Vo et al. (2015). Purified IgGs from the PMx-R2-2C antiserum

(Ndowora & Lockhart, 2000) were kindly provided by B. E. L. Lockhart, The University of Minnesota. To generate anti-peptide antisera, synthetic peptides that were conjugated to keyhole limpet haemocyanin (KLH) protein via an additional N-terminal cysteine residue and a maleimidocaproyl-N-hydroxysuccinimide linker were prepared by Mimotopes Pty. The peptides were mixed with adjuvant in equal ratios and injected into a rabbit via subcutaneous and intramuscular routes at days 0 and 14 (complete Freund's adjuvant), followed by another five boosts, 7 days apart, with incomplete Freund's adjuvant; serum samples were taken on day 64. Unless otherwise indicated, IgG antibodies from all antisera were purified using a Melon Gel IgG Purification Kit (ThermoFisher Scientific).

SDS-PAGE. SDS-PAGE was done using a NuPage Novex 4–12 % Bis-Tris pre-cast gel (ThermoFisher Scientific) as per the manufacturer's instructions.

Protein sequencing. *De novo* peptide and N-terminal sequencing were outsourced to Proteomics International. Protein bands were excised following separation on a NuPAGE 4–12 % Bis-Tris gel and staining with SimplyBlue Safe Stain (Life Technologies), destained in water, trypsin-digested and analysed on a 4800 MALDI TOF/TOF Analyzer following the method of Bringans et al. (2008). MASCOT sequence matching software (Matrix Science) was used to search the Ludwig Institute NR (non-redundant) protein sequence database. For *de novo* sequence analysis, peptide mass spectra were interpreted using PEAKS 4.5 (Bioinformatics Solutions). For N-terminal sequencing, proteins were electro-blotted onto an Immobilon-PSQ PVDF membrane (Millipore) using an iBlot Transfer Device. Automated Edman Degradation was done using a Model 494 Procise Protein/Peptide Sequencing System (Applied Biosystems).

Protein modelling. RaptorX (Källberg et al., 2012), PSIPRED (McGuffin et al., 2000), DISOPRED2 (Ward et al., 2004a, b) and NLStradamus (Nguyen Ba et al., 2009) were used to predict conserved domains, protein folds, intrinsically disordered regions and the presence of nuclear localization signals, respectively. Multiple secondary structure alignments were done using PROMALS3-D (Pei et al., 2008). Robetta (Kim et al., 2004) and Threadom (Xue et al., 2013) were used for full chain protein structural prediction and protein domain boundary prediction, respectively. Then, I-TASSER, a software package that uses a combinatorial approach of comparative modelling, LOMETS threading (model building based on solved protein structures) and *ab initio* modelling, was used to construct 3D models of the BSMYV CP (Roy et al., 2010; Roy & Zhang, 2001). Empirically determined epitopes were visualized on the 3D protein model using UCSF Chimera v1.7 (Pettersen et al., 2004). Continuous B-cell epitopes were predicted using BepiPred 1.0 (Larsen et al., 2006).

Pepscan assays. Libraries of overlapping peptides, each 16 aa residues long and offset by 4 aa residues, were designed to span the entire CP and VAP of BSMYV and the N termini of the CPs of BSOLV, BSGFV, BSIMV and BSCAV (Table S1). The peptides were synthesized by Mimotopes and each had a free acid (–OH) at the carboxyl terminus and a hydrophilic tetrapeptide spacer (SGSG) at the amino terminus that was coupled to a biotin molecule. The lyophilized peptides were dissolved and diluted to the working concentration in acetic acid and acetonitrile according to Mimotopes' PepSets Technical Note.

Epitopes were identified using the screening method of Heuzenroeder et al. (2008) with minor modifications. Streptavidin-coated ELISA plates (Mimotopes) were first washed three times with PBS containing 0.1 % Tween 20 (PBS-T) and blocked with 200 µl of 2 % BSA in PBS-T for 2 h at room temperature. Approximately 30 pmol of biotinylated peptide in PBS-T+0.1 % BSA was then added to each well and incubated for 1 h at room temperature with gentle shaking. After three washes with PBS-T, 100 µl of either 0.15 µg ml^{-1} of JVQ1 or 0.5 µg ml^{-1} of PMxR2-2C

IgGs was added and incubated overnight at 4 °C. The plates were then washed with PBS-T as before and bound antibody was detected by the addition of 100 μl goat anti-rabbit IgG alkaline phosphatase-conjugated antibody (Sigma-Aldrich) diluted 1 : 10 000 in PBS-T+0.1 % BSA. Finally, after three washes each with PBS-T and then PBS, 100 μl of 1 mg ml^{-1} p-nitrophenyl phosphate in 10 mM diethanolamine (pH 9.5) was added, incubated at room temperature and A_{405} measured using a Multiskan EX microplate reader (ThermoFisher Scientific). Pre-immune (naïve) IgGs from the JVQ1 rabbit were used as a negative control and each peptide set was tested in triplicate.

The significance of differences in A_{405} between seroconverted and pre-immune IgGs was assessed using a linear mixed effects model of the form:

$$y_{ijkl} - \mu + b_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijkl}$$

where y_{ijkl} are the observed $\log_e(A_{405})$ values; μ is the grand mean; b_i is the average effect of the ith plate (i=1, 2, 3) and is normally distributed with a mean of zero and variance of σ^2_b; α is the average effect of the jth peptide (j=1, …, 83); β_k is the effect of the IgG treatment (k=1, 2); $(\alpha\beta)_{jk}$ is the interactive effect of the jth peptide with the kth IgG treatment; and ε_{ijkl} is the experimental error, which is normally distributed with a mean of zero and a variance of $\sigma^2\varepsilon$. The above model was fitted using the 'nlme' package for R version 2.15 (Pinheiro et al., 2011).

The interactive effect $(\alpha\beta)_{jk}$ is the term that explains whether there is a significant difference in $\log_e$(absorbance) between pre-immune and post-immune sera for each of the 83 peptides. After undertaking the linear mixed effects model for the three sets of data, the interactions between the pre- and post-immune rabbit antisera and peptide factors were examined to see whether these terms were significantly different from zero.

Anti-peptide antibody assays. For PTA ELISA, virus minipreps of BSMYV-infected leaf tissue and of equivalent quantities of healthy leaf tissue were diluted in 0.05 M sodium carbonate (pH 9.6) to a concentration of 10 μg protein ml^{-1}, then 100 μl aliquots were added to a Maxisorp flat-bottom 96-well microtitre plate (Nunc) and incubated at room temperature for 2 h. Wells were blocked with 100 μl of 2 % skimmed milk in PBS for at least 2 h. Two-fold serial dilutions of anti-peptide IgGs were made in PBS-T, starting at 20 μg ml^{-1}, and 100 μl aliquots were added to each well and incubated for 2 h at room temperature, followed by addition of 100 μl of a 1 : 20 000 (v/v) dilution of alkaline phosphatase-conjugated goat anti-rabbit IgG (Sigma-Aldrich) in PBS-T-+0.5 % skimmed milk, for the same period. Antibody detection steps were as described previously for the PepScan assays. After each incubation step, the plate was washed three times for 3 min each with PBS-T. All samples were tested in triplicate and the threshold of detection considered to be twice the mean absorbance value of the healthy samples.

Western blots were done as described by Vo et al. (2015), except the blotted proteins were detected using antibody–alkaline phosphatase conjugates prepared using a Lightning-Link Alkaline Phosphatase Labelling kit (Innova Biosciences) as per the manufacturer's instructions. ISEM (immunosorbent electron microscopy) and antibody decoration were done as described by Vo et al. (2015).

To test the specificity of the anti-peptide antibodies, the peptides were resynthesized with a single cysteine residue at the N terminus and cross-linked to BSA (Sigma-Aldrich) using sulfo-SMCC (Thermo Scientific) as follows. A BSA solution at 10 mg ml^{-1} in conjugation buffer (PBS, 5 mM EDTA, pH 7.0) was mixed at a ratio of 1 : 5 (w/w) with sulfo-SMCC dissolved in DMSO. The mixture was incubated at 25 °C for 60 min and then dialysed in a 10 kDa bag overnight at 4 °C in conjugation buffer. One milligram of peptide was then dissolved in 125 μl conjugation buffer containing 5 mM TCEP (Tris(2-carboxyethyl)phosphine), the solution

adjusted to pH 7.0 and left at room temperature for 30 min. Then, 35 μl of the peptide solution was mixed with 200 μl of BSA-SMCC solution (4 mg ml^{-1}) and incubated at 25 °C for 4 h for conjugation. Finally, PBS solution was added to a final volume of 800 μl. SDS-PAGE and Western blots were done as described before, using a 10 μl aliquot of BSA–peptide conjugate for analysis.

DAS ELISAs of BSA–peptide conjugates were done using standard methods (Clark & Adams, 1977) and the JVQ1 antibodies at a concentration of 6 μg ml^{-1} for coating the wells and the aforementioned antibody–AP conjugates for detection.

ACKNOWLEDGEMENTS

REFERENCES

Bringans, S., Kendrick, T. S., Lui, J. & Lipscombe, R. (2008). A comparative study of the accuracy of several de novo sequencing software packages for datasets derived by matrix-assisted laser desorption/ionisation and electrospray. Rapid Commun Mass Spectrom **22**, 3450–3454.

Champagne, J., Benhamou, N. & Leclerc, D. (2004). Localization of the N-terminal domain of cauliflower mosaic virus coat protein precursor. Virology **324**, 257–262.

Chapdelaine, Y., Kirk, D., Karsies, A., Hohn, T. & Leclerc, D. (2002). Mutation of capsid protein phosphorylation sites abolishes cauliflower mosaic virus infectivity. J Virol **76**, 11748–11752.

Cheng, C. P., Lockhart, B. E. & Olszewski, N. E. (1996). The ORF I and II proteins of Commelina yellow mottle virus are virion-associated. Virology **223**, 263–271.

Clark, M. F. & Adams, A. N. (1977). Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. J Gen Virol **34**, 475–483.

Debouck, C., Gorniak, J. G., Strickler, J. E., Meek, T. D., Metcalf, B. W. & Rosenberg, M. (1987). Human immunodeficiency virus protease expressed in Escherichia coli exhibits autoprocessing and specific maturation of the gag precursor. Proc Natl Acad Sci U S A **84**, 8903–8906.

Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol **6**, 197–208.

Geering, A. D., Pooggin, M. M., Olszewski, N. E., Lockhart, B. E. & Thomas, J. E. (2005). Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. Arch Virol **150**, 787–796.

Geysen, H. M., Meloen, R. H. & Barteling, S. J. (1984). Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. Proc Natl Acad Sci U S A **81**, 3998–4002.

Guerra-Peraza, O., de Tapia, M., Hohn, T. & Hemmings-Mieszczak, M. (2000). Interaction of the cauliflower mosaic virus coat protein with the pregenomic RNA leader. J Virol **74**, 2067–2072.

Harper, G. & Hull, R. (1998). Cloning and sequence analysis of banana streak virus DNA. Virus Genes **17**, 271–278.

Harper, G., Hart, D., Moult, S., Hull, R., Geering, A. & Thomas, J. (2005). The diversity of Banana streak virus isolates in Uganda. Arch Virol **150**, 2407–2420.

Hay, J., Grieco, F., Druka, A., Pinner, M., Lee, S. C. & Hull, R. (1994). Detection of rice tungro bacilliform virus gene products *in vivo*. *Virology* **205**, 430–437.

Heuzenroeder, M. W., Barton, M. D., Vanniasinkam, T. & Phumoonna, T. (2009). Linear B-cell epitope mapping using enzyme-linked immunosorbent assay for libraries of overlapping synthetic peptides. In *Epitope Mapping Protocols*, 2nd edn, pp. 137–144. Edited by M. Schutkowski & U. Reineke. Totowa, NJ: Humana Press.

Hoh, F., Uzest, M., Drucker, M., Plisson-Chastang, C., Bron, P., Blanc, S. & Dumas, C. (2010). Structural insights into the molecular mechanisms of cauliflower mosaic virus transmission by its insect vector. *J Virol* **84**, 4706–4713.

Jacquot, E., Hagen, L. S., Jacquemond, M. & Yot, P. (1996). The open reading frame 2 product of cacao swollen shoot badnavirus is a nucleic acid-binding protein. *Virology* **225**, 191–195.

Jacquot, E., Keller, M. & Yot, P. (1997). A short basic domain supports a nucleic acid-binding activity in the rice tungro bacilliform virus open reading frame 2 product. *Virology* **239**, 352–359.

Karsies, A., Hohn, T. & Leclerc, D. (2001). Degradation signals within both terminal domains of the cauliflower mosaic virus capsid protein precursor. *Plant J* **27**, 335–343.

Karsies, A., Merkle, T., Szurek, B., Bonas, U., Hohn, T. & Leclerc, D. (2002). Regulated nuclear targeting of cauliflower mosaic virus. *J Gen Virol* **83**, 1783–1790.

Kim, D. E., Chivian, D. & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526–W531.

King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (2012). *Virus Taxonomy: Classification and Nomenclature of Viruses. Ninth Report of the International Committee on Taxonomy of Viruses.* London, UK: Elsevier.

Kobayashi, K., Tsuge, S., Stavolone, L. & Hohn, T. (2002). The cauliflower mosaic virus virion-associated protein is dispensable for viral replication in single cells. *J Virol* **76**, 9457–9464.

Kringelum, J. V., Nielsen, M., Padkjær, S. B. & Lund, O. (2013). Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol Immunol* **53**, 24–34.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H. & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* **7**, 1511–1522.

Laco, G. S. & Beachy, R. N. (1994). Rice tungro bacilliform virus encodes reverse transcriptase, DNA polymerase, and ribonuclease H activities. *Proc Natl Acad Sci U S A* **91**, 2654–2658.

Larsen, J. E., Lund, O. & Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res* **2**, 1–7.

Leclerc, D., Chapdelaine, Y. & Hohn, T. (1999). Nuclear targeting of the cauliflower mosaic virus coat protein. *J Virol* **73**, 553–560.

Marmey, P., Bothner, B., Jacquot, E., de Kochko, A., Ong, C. A., Yot, P., Siuzdak, G., Beachy, R. N. & Fauquet, C. M. (1999). Rice tungro bacilliform virus open reading frame 3 encodes a single 37-kDa coat protein. *Virology* **253**, 319–326.

Marmey, P., Rojas-Mendoza, A., de Kochko, A., Beachy, R. N. & Fauquet, C. M. (2005). Characterization of the protease domain of *Rice tungro bacilliform virus* responsible for the processing of the capsid protein from the polyprotein. *Virol J* **2**, 33.

Martinez-Izquierdo, J. & Hohn, T. (1987). Cauliflower mosaic virus coat protein is phosphorylated in vitro by a virion-associated protein kinase. *Proc Natl Acad Sci U S A* **84**, 1824–1828.

Martinière, A., Bak, A., Macia, J. L., Lautredou, N., Gargani, D., Doumayrou, J., Garzo, E., Moreno, A., Fereres, A. & other authors (2013). A virus responds instantly to the presence of the vector on the host and forms transmission morphs. *Elife* **2**.

McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405.

Mougeot, J. L., Guidasci, T., Wurch, T., Lebeurier, G. & Mesnard, J. M. (1993). Identification of C-terminal amino acid residues of cauliflower mosaic virus open reading frame III protein responsible for its DNA binding activity. *Proc Natl Acad Sci U S A* **90**, 1470–1473.

Ndowora, T. C. R. & Lockhart, B. E. L. (2000). Development of a serological assay for detecting serologically diverse banana streak virus isolates. *Acta Hortic* **540**, 377–388.

Nguyen Ba, A. N., Pogoutse, A., Provart, N. & Moses, A. M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* **10**, 202.

Pei, J., Tang, M. & Grishin, N. V. (2008). PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, W30–W34.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera– a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612.

Pinheiro, J., DebRoy, S., Sarkar, D. & R Development Core Team (2011). nlme: Linear and nonlinear mixed effects models. R package version 2.15. http://cran.rproject.org/web/packages/nlme/index.html.

Plisson, C., Uzest, M., Drucker, M., Froissart, R., Dumas, C., Conway, J., Thomas, D., Blanc, S. & Bron, P. (2005). Structure of the mature P3-virus particle complex of cauliflower mosaic virus revealed by cryo-electron microscopy. *J Mol Biol* **346**, 267–277.

Prabu-Jeyabalan, M., Nalivaika, E. & Schiffer, C. A. (2002). Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **10**, 369–381.

Qu, R. D., Bhattacharyya, M., Laco, G. S., De Kochko, A., Rao, B. L., Kaniewska, M. B., Elmer, J. S., Rochester, D. E., Smith, C. E. & Beachy, R. N. (1991). Characterization of the genome of rice tungro bacilliform virus: comparison with *Commelina* yellow mottle virus and caulimoviruses. *Virology* **185**, 354–364.

Roy, A. & Zhang, Y. (2001). Protein structure prediction. In *eLS*. Chichester: John Wiley & Sons, Ltd.

Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725–738.

Sandmeyer, S. B. & Clemens, K. A. (2010). Function of a retrotransposon nucleocapsid protein. *RNA Biol* **7**, 642–654.

Stavolone, L., Villani, M. E., Leclerc, D. & Hohn, T. (2005). A coiled-coil interaction mediates cauliflower mosaic virus cell-to-cell movement. *Proc Natl Acad Sci U S A* **102**, 6219–6224.

Takemoto, Y. & Hibi, T. (2001). Genes Ia, II, III, IV and V of Soybean chlorotic mottle virus are essential but the gene Ib product is non-essential for systemic infection. *J Gen Virol* **82**, 1481–1489.

Tompa, P. & Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**, 2–8.

Tözsér, J. (2010). Comparative studies on retroviral proteases: substrate specificity. *Viruses* **2**, 147–165.

Uversky, V. N. (2013). A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci* **22**, 693–724.

Vo, J. N., Mahfuz, N. N., Lockhart, B. E. L. & Geering, A. D. W. (2015). Improved methods for the purification and enrichment of banana streak virus for antibody production and protein detection. *Eur J Plant Pathol* **143**, 619–626.

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. (2004a). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004b). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635–645.

Xiong, Y. & Eickbush, T. H. (1990). Origin and evolution of retro-elements based upon their reverse transcriptase sequences. *EMBO J* **9**, 3353–3362.

Xue, Z., Xu, D., Wang, Y. & Zhang, Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **29**, i247–i256.