CrossMark

electronic reprint

BIOLOGICAL
CRYSTALLOGRAPHY

ISSN: 1399-0047
journals.iucr.org/d

Selecting soluble/foldable protein domains through single-gene or genomic ORF filtering: structure of the head domain of *Burkholderia pseudomallei* antigen BPSL2063

Louise J. Gourlay, Clelia Peano, Cecilia Deantonio, Lucia Perletti, Alessandro Pietrelli, Riccardo Villa, Elena Matterazzo, Patricia Lassaux, Claudio Santoro, Simone Puccio, Daniele Sblattero and Martino Bolognesi

Acta Cryst. (2015). D**71**, 2227–2235

IUCr Journals
CRYSTALLOGRAPHY JOURNALS ONLINE

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site or institutional repository provided that this cover page is retained. Republication of this article or its storage in electronic databases other than as specified above is not permitted without prior permission in writing from the IUCr.

For further information see http://journals.iucr.org/services/authorrights.html

BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Selecting soluble/foldable protein domains through single-gene or genomic ORF filtering: structure of the head domain of *Burkholderia pseudomallei* antigen BPSL2063

Louise J. Gourlay,[a]‡ Clelia Peano,[b]‡ Cecilia Deantonio,[c] Lucia Perletti,[a] Alessandro Pietrelli,[b] Riccardo Villa,[a] Elena Matterazzo,[a] Patricia Lassaux,[a] Claudio Santoro,[c] Simone Puccio,[b,d] Daniele Sblattero[e]* and Martino Bolognesi[a]*

[a]Department of Biosciences, University of Milan, Via Celoria 26, 20133 Milan, Italy, [b]Institute of Biomedical Technologies, National Research Council, Via Fratelli Cervi 93, 20090 Segrate, Italy, [c]Department of Health Sciences and IRCAD, University of Eastern Piedmont, Via Solaroli 17, 28100 Novara, Italy, [d]Doctoral School of Molecular and Translational Medicine, University of Milan, 20009 Segrate, Italy, and [e]Department of Life Sciences, University of Trieste, Via Weiss 2, 34128 Trieste, Italy. *Correspondence e-mail: dsblattero@units.it, martino.bolognesi@unimi.it

Received 2 July 2015
Accepted 21 August 2015

Edited by K. Miki, Kyoto University, Japan

‡ These authors should be considered joint first authors.

Keywords: *Burkholderia pseudomallei*; open reading frame-filtering library; protein antigen structure; soluble domain selection.

PDB reference: residues 657–992 of BPSL2063, 4usx

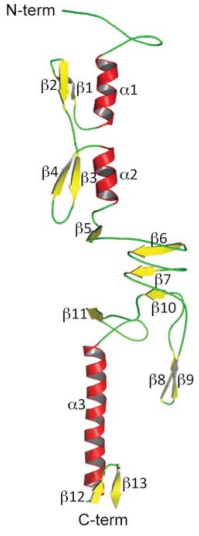
Supporting information: this article has supporting information at journals.iucr.org/d

The 1.8 Å resolution crystal structure of a conserved domain of the potential *Burkholderia pseudomallei* antigen and trimeric autotransporter BPSL2063 is presented as a structural vaccinology target for melioidosis vaccine development. Since BPSL2063 (1090 amino acids) hosts only one conserved domain, and the expression/purification of the full-length protein proved to be problematic, a domain-filtering library was generated using β-lactamase as a reporter gene to select further BPSL2063 domains. As a result, two domains (D1 and D2) were identified and produced in soluble form in *Escherichia coli*. Furthermore, as a general tool, a genomic open reading frame-filtering library from the *B. pseudomallei* genome was also constructed to facilitate the selection of domain boundaries from the entire ORFeome. Such an approach allowed the selection of three potential protein antigens that were also produced in soluble form. The results imply the further development of ORF-filtering methods as a tool in protein-based research to improve the selection and production of soluble proteins or domains for downstream applications such as X-ray crystallography.

1. Introduction

Melioidosis is a severe invasive disease caused by the Gram-negative saprotrophic bacterium *Burkholderia pseudomallei* (Cheng & Currie, 2005). Fatal clinical outcomes, coupled with poor control by antibiotics, have led to the search for potential vaccine components. In this context, a protein microarray probed with immune sera from melioidosis seropositive and recovery patients was formulated to identify potential antigens (Felgner *et al.*, 2009; Suwannasaen *et al.*, 2011). One seroreactive antigen identified by Felgner and coworkers is BPSL2063, a 1090-residue trimeric autotransporter adhesin (TAA; Felgner *et al.*, 2009), also termed BpaB (*Burkholderia pseudomallei* autotransporter B), shown to have a role in adhesion and invasion (Campos *et al.*, 2013). In fact, TAAs are known immunogenic components; notably, the neisserial adhesin A (NadA) TAA is one of the components of the recently developed serogroup B meningococcal vaccine 4CMenB (Felgner *et al.*, 2009; Giuliani *et al.*, 2006).

As their name suggests, TAAs are trimeric, display adhesion-like activities and, despite extensive sequence diversity, contain conserved structural elements, as shown by the three-



© 2015 International Union of Crystallography

electronic reprint

dimensional structures of several TAAs from diverse species (Dautin & Bernstein, 2007; Edwards *et al.*, 2010; Meng *et al.*, 2006, 2008; Nummelin *et al.*, 2004; Szczesny *et al.*, 2008). TAAs, like classical autotransporters, comprise an N-terminal passenger domain that can range from 200 to 3000 residues and a C-terminal β-domain that anchors the protein to the outer membrane and facilitates the transport of the passenger domain into the extracellular space (Dautin & Bernstein, 2007).

We are involved in a structural vaccinology project that focuses on the crystal structure analyses of *B. pseudomallei* seroreactive antigens as the foundations for computational biology-guided epitope-discovery optimization to develop immunogenic vaccine components (Felgner *et al.*, 2009). In this context, structure-based epitope-discovery optimization has already proved successful for other *B. pseudomallei* antigens (Gourlay *et al.*, 2013, 2015; Lassaux *et al.*, 2013; Malito & Rappuoli, 2013). Here, we report the high-resolution crystal structure analysis of the only conserved domain (residues 657–992) from BPSL2063 detectable based on sequence considerations.

As the crystallized domain ($BPSL2063_{Xtal}$) covers about one third of the full-length protein, and considering that epitopes may be present in other regions of the protein, we aimed to explore and design additional constructs covering the whole length of the protein. Unfortunately, it was not possible to produce the full-length form of BPSL2063 in bacterial cells and, owing to a lack of sequence identity with other TAAs, the design of additional domain constructs proved challenging. To address this issue, we applied our platform of open reading frame (ORF)-filtering analyses (D'Angelo *et al.*, 2011; Di Niro *et al.*, 2010; Zacchi *et al.*, 2003) to BPSL2063 using random fragmentation of the gene and TEM-1 β-lactamase as a folding reporter coupled to next-generation sequencing (NGS). After library sequencing, mapping reads help to select ORF fragments that correspond to protein domains that are likely to be soluble and correctly folded (D'Angelo *et al.*, 2011). For BPSL2063, the domain-filtered library identified two fragments corresponding to residues 99–434 ($BPSL2063_{D1}$) and residues 718–1037 ($BPSL2063_{D2}$), respectively, which were cloned and expressed in soluble form in *Escherichia coli* and purified for crystallization trials.

Finally, as a general tool for the identification of soluble proteins/domains for downstream applications, including three-dimensional structural studies, we report the construction of a *B. pseudomallei* genomic ORF-filtering library. This approach, *via* subsequent functional enrichment analyses, allowed us to identify three main protein functional categories that were over-represented among the genes: cell-envelope biogenesis, inorganic transport and nucleotide transport and metabolism. Among these, three antigen genes (BPSL1801, BPSL1626 and BPSL2520) that are the focus of our structural vaccinology project were represented in the library; using the boundaries identified by our filtering tool, constructs were successfully designed leading to soluble proteins for further applications. In fact, the crystal structure of one of these antigens (BPSL2520) has been solved (to be published).

Our methods and results, although preliminary, imply that a domain-filtering library approach may assist in the selection of protein domain boundaries and may accelerate and improve the steps leading to soluble protein production for downstream biochemical studies, such as crystallogenesis, where protein stability and solubility are key issues.

2. Materials and methods

2.1. BPSL2063 domain-filtering library construction

The BPSL2063 gene encoding for amino acids 55–1090 (lacking the first 54 N-terminal residues corresponding to the extended signal peptide region) was chemically synthesized and cloned into pUC57 vector (GenScript). To facilitate subsequent cloning, NdeI and BamHI restriction sites were added to the 5′ and 3′ ends of the gene, respectively. Following digestion, the fragment of 3111 nucleotides was collected. Total BPSL2063 DNA (2 µg) was fragmented using a bench sonicator (Covaris S2) with conditions optimized to obtain a range of 200–800 bp fragments (duty cycle, 10%; intensity, 5.0; cycles per burst, 200; duration, 2 × 60 s, total 120 s; mode, frequency sweeping; temperature, 6°C). BPSL2063 DNA fragments were purified, blunt-ended using the Quick Blunting Kit (New England Biosciences) and cloned into pFILTER vector (D'Angelo *et al.*, 2011). After ligation, the BPSL2063 library was electroporated into *E. coli* BL21(DE3) cells and plated on chloramphenicol plates supplemented with ampicillin at the highest concentration (100 µg ml^{-1}). 10^6 bacterial cells were plated, and after overnight growth 10^4 colonies were obtained.

2.2. *B. pseudomallei* genomic ORF-filtering library construction

B. pseudomallei strain K96243 genomic DNA (gDNA; kindly supplied by Professor R. Titball, University of Exeter, England) was fragmented by ultrasonication (Covaris S2; duty cycle, 10%; intensity, 5.0; cycles per burst, 200; duration 2 × 60 s, total 120 s; mode, frequency sweeping; temperature, 6°C); a range of 200–800 bp fragments was obtained. Fragments were collected, purified, blunt-ended using the Quick Blunting Kit (New England Biosciences) and cloned into pFILTER vector using an EcoRV cloning site between a pelB leader sequence and the mature β-lactamase gene (D'Angelo *et al.*, 2011). After ligation, the *B. pseudomallei* genomic DNA library was electroporated into *E. coli* BL21(DE3) cells and plated on chloramphenicol plates supplemented with ampicillin at the highest concentration (100 µg ml^{-1}). 5×10^7 bacterial cells were plated, and after overnight growth 10^5 colonies were obtained.

2.3. ORF- and domain-filtering library sequencing and analysis

ORF fragments were excised from the pFILTER vector by double digestion (BssHII and NheI) and 200 ng of DNA fragments were purified using MinElute columns (Qiagen) in order to remove shorter fragments. Ligation of the purified

electronic reprint

Table 1
The cloning regions of the six genes/domains are illustrated in comparison with the fragments that were represented in the *B. pseudomallei* genomic ORF-filtering library.

Cloned residues and numbers in parentheses refer to the amino-acid sequence.

Gene	Mapped region	Cloned residues
BPSL1626	1883679 (1)–1884209 (176)	26–176
BPSL1801	2143929 (31)–2144346 (170)	21–171
BPSL2520	3035865 (9)–3036434 (198)	21–198
$BPSL2063_{D1}$	2468275 (99)–2469280 (434)	99–434
$BPSL2063_{D2}$	2470132 (718)–2471091 (1037)	712–1037
$BPSL2063_{Xtal}$	—	657–992

samples to specific adaptors and preparation of the sequencing libraries was performed following the Rapid Library preparation-method manual (Roche, 454 Titanium). DNA libraries were quantified using a PicoGreen DNA Quantitation Kit (Invitrogen) and checked for quality by capillary electrophoresis (Agilent Bioanalyzer 2100 with High Sensitivity DNA assay kit; Agilent Technologies). DNA libraries were then amplified in emulsion following the Titanium LIB-L em-PCR protocol (Roche). Reactions were recovered by 2-propanol emulsion breaking, and beads presenting clonally amplified DNA fragments on their surface were enriched. Each enriched sample was separately loaded onto one-eighth of a PicoTiterPlate (PTP) and sequenced according to the 454 GS-FLX Titanium protocol.

In the first data-analysis pipeline, raw reads from both the sequencing of the genomic ORF-filtering library and of the domain-filtered BPSL2063 library were mapped against the *B. pseudomallei* K96243 reference genome using the CLC Genomics Workbench with default parameters. CLC mapping files were exported as BAM files and the *BEDTools* package was used to calculate the statistics for depth and coverage (Quinlan & Hall, 2010). We performed the functional enrichment test using the *R* software package to compute Fisher test statistics (R Development Core Team, 2011).

In the second data-analysis pipeline, raw reads from the sequencing of the genomic ORF-filtering library were analysed using the recently developed *NGS-TreX* web interface (http://www.ngs-trex.org), which has previously been applied to the analysis of ORF-filtered genomic libraries (D'Angelo *et al.*, 2013; Boria *et al.*, 2013).

2.4. Cloning, expression and purification of $BPSL2063_{D1}$, $BPSL2063_{D2}$ and $BPSL2063_{Xtal}$

$BPSL2063_{D1}$ (residues 99–434), $BPSL2063_{D2}$ (residues 712–1037) and $BPSL2063_{Xtal}$ (residues 657–992) were amplified from the chemically synthesized BPSL2063 gene cloned into the pUC57 vector (GenScript) by PCR amplification using Phusion DNA polymerase (Thermo Scientific) according to standard protocols. For $BPSL2063_{D1}$ and $BPSL2063_{D2}$, primers were designed according to pET151-D-TOPO requirements (Life Technologies). For $BPSL2063_{Xtal}$, primers included BamHI and EcoRI restriction sites for cloning into pET-21b (Novagen). Cloning regions are given in Table 1 and primer details are reported in Supplementary Table S1. Successful cloning and PCR fidelity were confirmed by DNA sequencing (Eurofins Genomics).

The three domains were expressed as His-tag fusion proteins in *E. coli* BL21(DE3) Star cells (Invitrogen) or Tuner (DE3) cells (Novagen) in LB broth, inducing with 0.1 m*M* IPTG at 18°C overnight. Initial solubility tests were carried out on 10 ml bacterial cultures using BugBuster Protein Extraction Reagent (Merck) according to the maufacturer's instructions. Large-scale expression was carried out for $BPSL2063_{D2}$ and $BPSL2063_{Xtal}$. Bacterial cells from a 0.25 l culture were harvested and lysed in IMAC wash buffer 1 (see above) containing one tablet of cOmplete Protease Inhibitor Cocktail (Roche), lysozyme (0.20 mg ml^{-1}), DNases (50 µg ml^{-1}) and 10 m*M* $MgCl_2$. Cell lysis and purification were as reported in the previous section. $BPSL2063_{D2}$ was further purified by size-exclusion chromatography on a Superdex 200 (60/600) gel-filtration column (GE Healthcare) pre-equilibrated with crystallization buffer (10 m*M* Tris pH 8.0). $BPSL2063_{Xtal}$ was exchanged into the same buffer using a PD-10 desalting column (GE Healthcare). Protein purity was judged by SDS–PAGE on a 12% NuPAGE Bis-Tris Precast Gel with NuPAGE MES SDS Running Buffer (Life Technologies). Both protein samples were concentrated using Amicon Ultra Centrifugal Filters (Millipore) for crystallization trials.

2.5. Crystallization of $BPSL2063_{Xtal}$

Crystallization trials for $BPSL2063_{D2}$ (11.6 mg ml^{-1}) and $BPSL2063_{Xtal}$ (3.5 mg ml^{-1}) were set up in 96-well Greiner sitting-drop plates containing 100 µl crystallization solution using an Oryx 8 robot (Douglas Instruments). Crystals of $BPSL2063_{Xtal}$ grew overnight at 20°C in a 400 nl drop (70% protein) in condition D1 [0.1 *M* malic acid, MES and Tris (MMT) buffer pH 4.0, 25% PEG 1500] of the PACT Premier screen (Molecular Dimensions). Crystals were immersed in mother liquor containing an elevated concentration of PEG 1500 (40%) and flash-cooled in liquid nitrogen. Crystallization experiments on $BPSL2063_{D2}$ are ongoing; however, in light of the proven necessity of proteolysis for the successful crystallization of both $BPSL2063_{Xtal}$ and BpaA (Edwards *et al.*, 2010) as discussed later, controlled proteolysis of $BPSL2063_{D2}$ may be necessary.

2.6. $BPSL2063_{Xtal}$ data collection, structure determination and refinement

Diffraction data were collected from a single crystal to 1.8 Å resolution (on beamline ID23-2 at the European Synchrotron Radiation Facility, Grenoble, France; Supplementary Table S3). Data were processed using *XDS*, assigned to the monoclinic space group $P2_1$ and scaled using *POINTLESS* and *SCALA*, respectively, in the *CCP*4 suite (Evans, 2006; Kabsch, 2010). The three-dimensional structure of $BPSL2063_{Xtal}$ was solved by molecular replacement with *MOLREP* (Vagin & Teplyakov, 2010) using the structure of the head domain of a TAA from *B. pseudomallei* strain 1710b

electronic reprint

(BpaA; PDB entry 3laa) as a search model (67% sequence identity over 153 residues; Edwards *et al.*, 2010). The structure was fitted to the generated electron-density maps using *Coot* and was further refined using *phenix.refine* (Emsley & Cowtan, 2004; Afonine *et al.*, 2012). All data were refined to satisfactory final R_{gen} and R_{free} factors of 16.8 and 20.4%, respectively, and the stereochemical parameters were checked using *MolProbity* in the *PHENIX* platform (Supplementary Table S4; Chen *et al.*, 2010; Davis *et al.*, 2007). Atomic coordinates and structure factors have been deposited in the Protein Data Bank (Berman *et al.*, 2000) as PDB entry 4usx.

2.7. Cloning, expression and purification of BPSL1801, BPSL1626 and BPSL2520

BPSL1801, BPSL1626 and BPSL2520, identified through the *B. pseudomallei* genomic ORF-filtering library, were cloned as N-terminal histidine-tag fusion proteins (Supplementary Table S1) and expressed in *E. coli* BL21(DE3) Star (BPSL1626), Rosetta (DE3)/pLysS (BPSL1801) or C41(DE3)/pLysS (BPSL2520) cells in Luria–Bertani (LB) broth, inducing with 0.5 m*M* IPTG, at 20°C overnight. All fusion proteins were purified from 0.5 l bacterial cultures, harvested, resuspended in IMAC wash buffer 1 (50 m*M* potassium dihydrogen phosphate pH 8.0, 0.3 *M* potassium chloride, 5 m*M* imidazole) containing lysozyme (0.20 mg ml^{-1}), DNases (50 µg ml^{-1}) and 10 m*M* magnesium chloride. Bacteria were lysed using a Cell Disruption System (Constant Systems) at 25 MPa. The soluble fraction was obtained by centrifugation at 16 000 rev min^{-1} for 20 min and was loaded onto a 5 ml Bio-Scale Mini Profinia IMAC cartridge (Bio-Rad) pre-equilibrated with IMAC wash buffer 1. Proteins were washed with IMAC wash buffer 2 containing 10 m*M* imidazole and were eluted with IMAC elution buffer containing 250 m*M* imidazole.

3. Results

3.1. Primary structure analysis of BPSL2063

BPSL2063 (UniProt code Q63TA4) comprises 1090 amino acids and is one of nine proteins belonging to the TAA protein family involved in adhesion to host cells (Campos *et al.*, 2013; Lazar Adler *et al.*, 2011). BPSL2063 contains an extended signal peptide (ESPR; residues 1–54) typical of proteins secreted by the type V secretion system, including TAAs (also termed type Vc secretion-system proteins; Desvaux *et al.*, 2006; Henderson *et al.*, 1998). BPSL2063 shares several short conserved elements common to the first TAA characterized, the *Yersinia enterocolitica* adhesin A (YadA): there are five YadA stalk regions (PFAM PF05662) located between residues 478 and 989, a YadA-like head region (PF05658; residues 925–951) and a YadA C-terminal membrane-anchoring domain (PF03895; residues 1015–1090).

3.2. Three-dimensional structure analyses of BPSL2063$_{Xtal}$

BPSL2063$_{Xtal}$ was expressed, purified and crystallized as described in §2 (Fig. 1, Supplementary Tables S1 and S2). The 1.8 Å resolution crystal structure of BPSL2063$_{Xtal}$ was solved by molecular replacement using the structure of BpaA (PDB entry 3laa; Edwards *et al.*, 2010) as the search model and was refined to convergence (R_{gen} = 16.8%, R_{free} = 20.4%; Figs. 2*a* and 2*b* and Supplementary Tables S3 and S4; see §2). There are three BPSL2063$_{Xtal}$ chains (*A*, *B* and *C*; r.m.s.d. of 0.6–0.68 Å) present in the crystal asymmetric unit which interlock to form the biologically relevant trimer (Figs. 2*a* and 2*b*). Trimerization of TAAs is in fact essential for their translocation and function. Electron density was visible for residues 678–891, 678–890 and 677–892 in chains *A*, *B* and *C*, respectively, corresponding to the absence of about 119 residues at the C-terminus (19 residues derived from the vector),

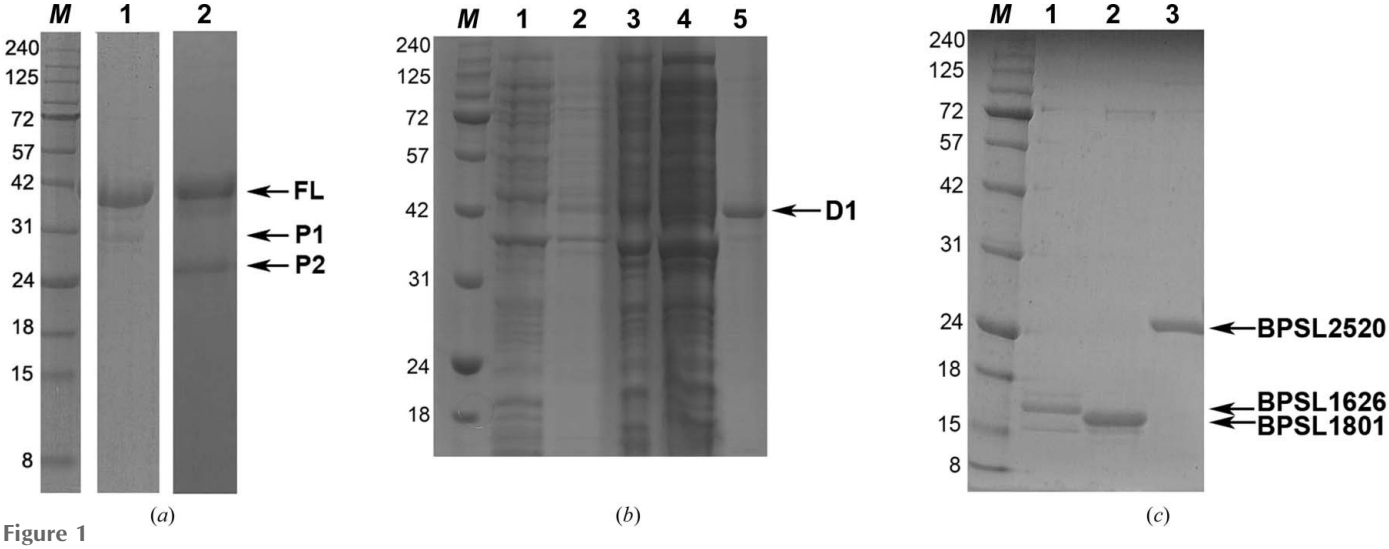


Figure 1
SDS–PAGE analysis of recombinant proteins. (*a*) Purified BPSL2063$_{Xtal}$ (lane 1) and BPSL2063$_{D2}$ (lane 2). The top arrow indicates the full-length (FL) form of the domains, whereas arrows P1 and P2 indicate two lower molecular-weight fragments observed for BPSL2063$_{Xtal}$ and BPSL2063$_{D2}$, respectively. (*b*) Expression and solubility tests of BPSL2063$_{D1}$ showing the non-induced and induced cells (lanes 1 and 2, respectively), the total bacterial extract (lane 3) and the soluble (lane 4) and insoluble (lane 5) fractions following chemical cell lysis and centrifugation. (*c*) Purified BPSL1626 (lane 1), BPSL1801 (lane 2) and BPSL2520 (lane 3). In all panels, molecular-weight markers (Genespin) are indicated on the left in kDa.

electronic reprint

suggesting that BPSL2063$_{Xtal}$ had been subjected to proteolysis during, and possibly prior to, crystallization. This is consistent with an unlikely Matthews coefficient estimation for a trimer of the entire construct (371 residues, of which 35 correspond to vector regions), and with SDS–PAGE analysis of the purified protein, which highlighted a band at approximately 28 kDa (Fig. 1*a*). Controlled proteolysis was in fact found to be essential for the crystallization of BpaA (Edwards *et al.*, 2010). Electron density was also absent for the first 33–34 (16 vector residues) N-terminal residues (depending on the chain), indicating conformational flexibility in this region.

3.3. The three-dimensional BPSL2063$_{Xtal}$ fold

Following structure-based homology analyses using the *PDBeFold* server (http://www.ebi.ac.uk/msd-srv/ssm/), as expected, the BPSL2063$_{Xtal}$ chain was found to be structurally closely related to two other TAAs: BpaA (the search model; r.m.s.d. of 0.92 Å; 60.9% sequence identity) and a fragment (K9; residues 1049–1304) from the *Salmonella enterica* TAA SadA (PDB entry 2yo1; r.m.s.d. of 1.35 Å; 33.2% sequence identity; Edwards *et al.*, 2010; Hartmann *et al.*, 2012). (See Fig. 2*c* for a structural comparison of BPSL2063$_{Xtal}$ with BpaA.)

BPSL2063$_{Xtal}$ contains two YadA stalk regions (residues 682–699 and 829–852) and a YadA-like head region (residues 765–792; not predicted by PFAM), which forms a left-handed parallel β-roll. Structural comparisons between BPSL2063$_{Xtal}$ and BpaA were made (Fig. 2*c*). In contrast to the N-terminus of BpaA, which commences with a coiled coil, each BPSL2063$_{Xtal}$ monomer begins with a so-called 'neck' region (from the N-terminal residue to Asn695) that immediately precedes the first YadA-like stalk. According to domain annotation of the TAA nomenclature using the Bioinformatics Toolkit available from the Max Planck Institute (http://toolkit.tuebingen.mpg.de/dataa/), this region typically connects β-structured domains to coiled coils in the N- to C-terminal direction only (Szczesny & Lupas, 2008). The BPSL2063$_{Xtal}$ neck region is stabilized by several intermolecular hydrogen bonds and a salt bridge formed between His682 ND1 in one chain and Asp692 OD1 and OD2 in the adjacent chain.

The neck is followed by a α-helix–β-turn and β-turn–α-helix motif. α-Helix 1 partially superimposes with α-helix 1 from BpaA and is followed by a β-turn (β1–β2) absent in BpaA that is in turn followed by a second β-turn (β3–β4) present in BpaA and α-helix 2 (Fig. 2*c*). α-Helix 2 is followed by β-strand 5 that represents the first strand in a five-stranded β-sheet composed of three β-strands (β6, β7 and β10) from the adjacent chain and a fifth strand (β11) from the same monomer. This region encompasses the so-called YadA-like head motif (β-strands 3 and 4; residues 765–792) (Figs. 2*a* and 2*b*). Between β7 and β10, the polypeptide chain deviates and extends across the back of the trimer to form a β-turn (β8–β9). The polypeptide deviates to form a second neck connector domain and a C-terminal α-helix (α3). α3 contributed by each chain intertwine to form a left-handed coiled coil that contains a leucine zipper formed between Leu852, Leu862 and Leu869 and their equivalent residues in the other chains. There is a short gap in the electron density after residue 875 or 878 (chain *C*), followed by a β-turn (β12–β13) in chains *A* and *C* (non-structured in chain *B*) that build the C-terminus.

Several sequence motifs common to TAA members are present in BPSL2063$_{Xtal}$, namely (i) the NSVAIG TAA structural motif present in the YadA head domain (residues 765–792) of BPSL2063$_{Xtal}$ that mediates collagen

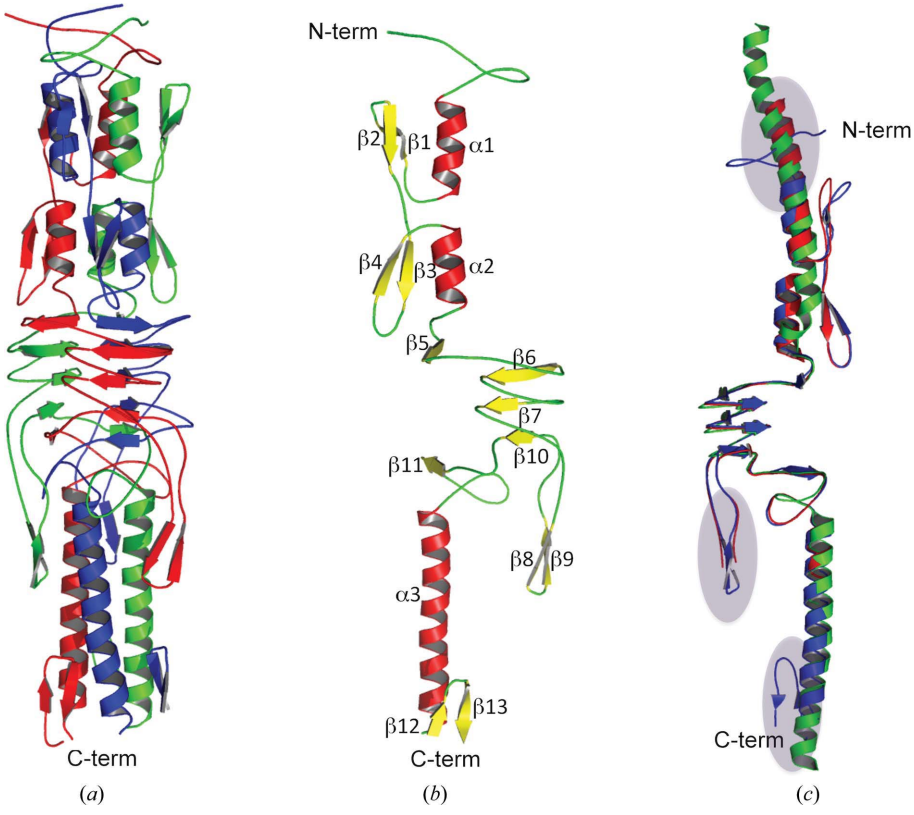


Figure 2
Three-dimensional structure of the BPSL2063$_{Xtal}$ trimer. (*a*) Secondary-structure cartoon representation of the three BPSL2063$_{Xtal}$ chains present in the crystal asymmetric unit. Chains *A*, *B* and *C* are shown in blue, green and red, respectively. The N- and C-termini are labelled. (*b*) Secondary-structure cartoon representation of BPSL2063$_{Xtal}$ chain *A* coloured according to secondary-structure element, with α-helices, β-strands and unstructured regions coloured red, yellow and green, respectively. (*c*) Secondary-structure cartoon representation of chain *A* from BpaA (red), residues 1180–1336 of chain *A* from the K9 fragment of SadA (green; PDB entry 1y01) and chain *C* from BPSL2063$_{Xtal}$ (blue), indicating the N- and C-termini of BPSL2063$_{Xtal}$ and the three main regions of structural diversity in shaded ovals. All panels were generated using *PyMOL*.

electronic reprint

binding in YadA (Nummelin *et al.*, 2004), (ii) a HANS connector motif, (iii) an FGG motif that is typically followed by a GA*x*Y motif, although $BPSL2063_{Xtal}$ does not conform to this signature, and (iv) a HIN2 (F*x*G) motif. The first structural representation of the HIN2 motif was proposed in the BpaA structure, although a five-residue gap in the electron density was observed (Edwards *et al.*, 2010). In $BPSL2063_{Xtal}$ the electron density is continuous and a β-turn (comprising β-strands 8 and 9) is present. Based on structural analyses of the $BPSL2063_{Xtal}$ structure, we may deduce that such structural differences are a consequence of how the BpaA construct was designed. As α3 in BpaA is truncated, this results in loss of the hydrogen bonds that stabilize this β-turn. Therefore, we propose the HIN2 motif from $BPSL2063_{Xtal}$ as the real first representative structure of this motif.

3.4. Construction of a BPSL2063 domain-filtering library and sequence analysis

Given the inability to express full-length BPSL2063 within the 1090-residue protein, we constructed a domain-filtering library (see §2) to identify potential soluble domains. 454 GS-FLX Titanium sequencing was performed on filtered fragments, and 5922 mapping reads with an average length of 294 bp for a total of 1 746 066 nucleotides (corresponding to a sequencing depth of 533×) were obtained. Despite the elevated sequencing depth, only two regions of the BPSL2063 gene were covered by mapping reads (base pairs 2 468 275 to 2 469 280 and 2 470 132 to 2 471 091), corresponding to amino-acid residues 99–434 ($BPSL2063_{D1}$) and 718–1037 ($BPSL2063_{D2}$), respectively. The latter construct extends over the $BPSL2063_{Xtal}$ cloning region; however, it is longer both at the N- and C-termini.

3.5. Production of $BPSL2063_{D1}$ and $BPSL2063_{D2}$

The two regions encoding for $BPSL2063_{D1}$ and $BPSL2063_{D2}$ were amplified from gDNA and cloned into pET151/TOPO and expressed as N-terminally His-tagged proteins in *E. coli* Tuner (DE3) cells ($BPSL2063_{D1}$) or BL21(DE3) Star cells ($BPSL2063_{D2}$), as described in §2. For $BPSL2063_{D2}$, it was necessary to clone from residue 712 instead of 718 owing to the excessive GC content (75%) of the nucleotide stretch that surrounds residue 718, which was undesirable for PCR amplification. The solubility of $BPSL2063_{D1}$ (Fig. 1*b*) was evidently lower than that of $BPSL2063_{D2}$, in accordance with the fact that the latter was represented to a greater extent in the domain-filtering library. Protein corresponding to both constructs was purified from the soluble fraction in yields of 1–2 mg ($BPSL2063_{D1}$) and 8 mg ($BPSL2063_{D2}$) per litre of bacterial culture (Fig. 1*a*). The purity of $BPSL2063_{D1}$ was poor and further attempts to improve sample purity (*e.g.* size-exclusion chromatography or the addition of detergent) resulted in insufficient yields; optimization of the purification procedure will be necessary for crystallization screening. $BPSL2063_{D2}$ was exchanged into 10 m*M* Tris–HCl pH 8.0 and concentrated for crystallization trials (see §2).

3.6. Construction of a *B. pseudomallei* genome ORF-filtering library

In order to extend our search methods to allow the identification of potential soluble domains on a larger scale, we applied our strategy based on the construction of an ORFeome library from the *B. pseudomallei* genome. By fragmenting a whole (intronless) genome into DNA fragments of 200–800 bp (D'Angelo *et al.*, 2011; Heger & Holm, 2003) it is possible to create a library of fragments coding for potential domains (or parts thereof). DNA fragments encoding well folded protein domains, fused upstream of β-lactamase, allow the reporter enzyme to fold correctly and allow bacteria to survive the selective pressure posed by the antibiotic. With this aim, *B. pseudomallei* strain K96243 gDNA was fragmented, cloned and filtered by plating onto plates containing increasing ampicillin concentrations. 10^7 bacterial cells were plated and, after overnight growth, approximately 1% of the cells survived at the highest ampicillin concentration, yielding a library with an estimated size of 3×10^5 (Fig. 3). The library size is estimated assuming an unbiased selection of real ORF fragments of 200–800 bp in length. Nevertheless, the real fragment distribution is derived from a combination of additional factors, including the possibility of selecting some spurious ORFs as well as the negative selection of some real ORFs owing to the inability of their translated products to fold.

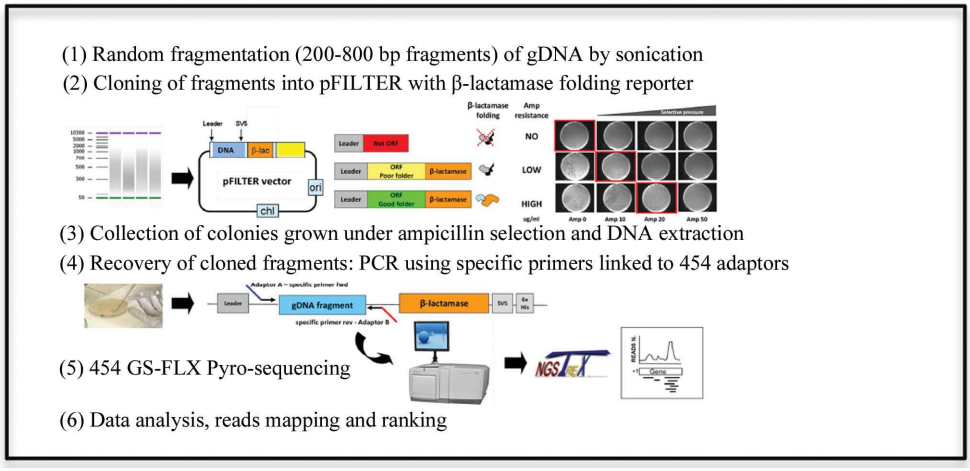


Figure 3
Schematic overview of the main steps in the construction of the *B. pseudomallei* genomic ORF-filtering library. 1, random fragmentation of genomic DNA. 2, gDNA fragment cloning into the pFILTER vector and filtering using β-lactamase as a folding reporter. 3, harvesting of *E. coli* transformants and DNA extraction. 4, gDNA fragment recovery by amplification using specific primers linked to adaptors for sequencing. 5, deep sequencing. 6, data analysis.

electronic reprint

Table 2
A list of the most represented ORFs in the *B. pseudomallei* genomic ORF-filtered library belonging to the three most enriched functional categories (depth of >10 for both pipelines and focus >0.8 with the *NGS-TreX* pipeline).

Locus ID	Strand	Gene coverage	Mean depth	Focus	Peak start	Peak end	Gene description
BPSL0466	+	1	54.3301	0.896	503530	504348	ABC-type metal ion-transport system, periplasmic component/surface antigen
BPSL0882	−	0.248963	25.6933	0.889	1023539	1023839	Chromate-transport protein ChrA
BPSL2351	+	0.189685	16.2189	0.941	2843904	2844338	Nitric oxide reductase large subunit
BPSL0824	−	0.008	14	0.917	958876	958883	ABC-type metal ion-transport system, periplasmic component/surface adhesin
BPSL3404	−	0.338836	12.7974	0.941	4042163	4042553	Na^+/H^+-antiporter NhaD and related arsenite permeases
BPSS0357	+	1	6146.87	0.994	499290	499841	Uncharacterized protein probably involved in high-affinity Fe^{2+} transport
BPSS0404	+	0.194891	22.8577	0.857	547293	547560	Cytochrome *c* peroxidase
BPSL1866	+	0.213035	31.242	0.824	2219853	2220072	Dihydroorotate dehydrogenase
BPSL2199	−	0.850214	48.4648	0.856	2640323	2640919	Cell wall-associated hydrolases (invasion-associated proteins)
BPSS0418	+	0.133936	99.2331	0.984	578817	578980	Capsule polysaccharide export protein

In order to fully classify ORF fragments present in the surviving bacteria, they were rescued from filtering plates and plasmid DNA was obtained. The inserts were cleaved by restriction-enzyme digestion and subjected to Roche 454 FLX Titanium sequencing as described in §2 (Fig. 3).

Reads were analyzed using two different data-analysis pipelines: the first is highly conservative and allows the detection of multiple highly represented domains in single proteins; the second is less conservative and identifies single domains with high sequencing depth covered by highly focused reads. Based on the first pipeline analysis, sequences were aligned with the reference *B. pseudomallei* genome (accession Nos. NC_006351.1 and NC_006350.1) and a total of 96 231 mapping reads with an average length of 225 bp were obtained, corresponding to 19 420 180 nucleotides. The number of genes covered by mapping reads was 739 (chromosome 1) and 540 (chromosome 2), with a total of 1279 ORFs being represented in the library (21.5% of the total BPS genes; see Supplementary Tables S5 and S6). According to the second data-analysis pipeline, *NGS-TreX*, a total of 1441 ORFs were covered by mapping reads, 161 of which had a sequencing depth higher than 10 and only 67 of which had a focus above 0.8 (see §2).

3.7. Functional enrichment analysis of genes represented in the genomic ORF-filtering library

Although a sequencing depth of about 4× for the coding portion of the genome was obtained, only one fifth of the ORFs were covered by mapping reads and considered as represented in the ORF-filtering library. On these bases, we extended our analysis to determine whether the filtering vector in this case had exerted a selection based not only on the folding characteristics of the domains but also on features deriving from ORF function or cellular localization. With this aim, we performed a functional enrichment analysis based on a Fisher test, and observed that the main functional categories that were over-represented among the genes filtered out by both pipelines were cell-envelope biogenesis (p-value ≤ 0.01); inorganic transport (p-value ≤ 0.01) and nucleotide transport and metabolism (p-value ≤ 0.05) (Supplementary Table S7). Our results imply that for complex bacterial genomes such as that of *B. pseudomallei* (two chromosomes, 7.2 Mb and 68% GC content) the ORF-filtering vector favours the selection of ORFs encoding for proteins that are potentially exposed on the cell surface, such as those involved in transport mechanisms or located in the cell membrane. Thus, the extension of the filtering power of our selection strategy to identify protein domains that are located in the outer membrane may be useful in the context of host–pathogen interaction studies and may prompt the identification of antigens. A list of the most represented ORFs in the genomic filtering library (depth >10 with both pipelines and focus >0.8 with the *NGS-TreX* pipeline), belonging to the three most enriched functional categories, is reported in Table 2. These genes and gene regions covered by mapping reads can be considered to be good candidates for further recombinant production and biochemical analyses.

Proteins involved in transport and in cell-envelope biogenesis often exhibit large, complex structures comprising highly hydrophobic intermembrane regions together with extracellular and intracellular domains. 5% of chromosome 1 genes and 7.2% of chromosome 2 genes represented in the ORF-filtered library and covered by sequencing were longer than 3000 bases, suggesting that they encode proteins larger than 1000 amino acids. Given the lack of functional annotations, construct design for recombinant protein production may prove to be challenging and to be the rate-limiting step for subsequent biochemical/structural analyses.

3.8. ORF-filtering library-guided antigen-construct design and recombinant production

As previously mentioned, in the context of a structural vaccinology project we are focusing on the crystal structure analysis of potential antigens to be used for *in silico*-based epitope-discovery optimization. A comparison of our list of potential antigens [previously selected *via* bioinformatics, transcriptomics (Professor X. Daura's group, Universitat Autònoma de Barcelona, Spain) or protein microarrays (Felgner *et al.*, 2009)] with the list of genes represented in the genomic ORF-filtered library enabled the design of constructs for BPSL1626, BPSL1801 and BPSL2520. We selected three genes having three different levels of mean sequencing depth

electronic reprint

but very high levels of gene coverage (>0.80). BPSL1626 is the second top gene listed, with a very high mean sequencing depth of 342.3 and a coverage of 1; BPSL1801 has a medium mean sequencing depth of 29.8 and a coverage of 0.81, and BPSL2520 has a low mean sequencing depth of 3 and a coverage of 0.95 (Supplementary Table S5; see Supplementary Fig. S1 for a plot of the reads mapped on the three selected genes). Fragments identified by the genomic ORF-filtered library corresponded to the full-length proteins, except for BPSL1801 (from residue 31; Table 1). Signal peptides in BPSL1626 and BPSL2520 were removed from the constructs as being unfavourable for crystallization.

All proteins were successfully produced in soluble form as N-terminally His-tagged species in *E. coli* (see §2). The yields of soluble purified protein were approximately 20, 5 and 100 mg per litre of bacterial culture for BPSL1626, BPSL1801 and BPSL2520, respectively (Fig. 1*c*). BPSL2520 readily yielded diffraction-quality crystals and its three-dimensional structure has been solved *via* molecular-replacement methods (to be published); BPSL1626 and BPSL1801 are undergoing crystallization trials.

4. Discussion

In the search for suitable components for a melioidosis vaccine, we have been exploring the surfaceome of *B. pseudomallei* for the three-dimensional structures of potential antigens. Here, we present the 1.8 Å resolution crystal structure of a conserved YadA-like domain from the potential antigen BPSL2063, which was recognized by antibodies from patients that tested positive for *B. pseudomallei* infection (Felgner *et al.*, 2009). The crystal structure will serve as the basis for the application of computational epitope-prediction methods, which have already proved successful for other *B. pseudomallei* antigens identified by Felgner and coworkers (Felgner *et al.*, 2009; Gaudesi *et al.*, 2015; Gourlay *et al.*, 2013, 2015; Lassaux *et al.*, 2013). Although sequence conservation is limited among TAA members, they share three-dimensional structural features. The overall fold of $BPSL2063_{Xtal}$ reported here reflects that of other TAA members, displaying the canonical elongated and tightly intertwined trimeric assembly, and comprises classical TAA modules such as the YadA head and neck motifs. It also displays conserved short sequence motifs, such as the HIN2 motif with its F*x*G signature sequence. With regard to the latter motif, $BPSL2063_{Xtal}$ presents a continuous model in contrast to the five-residue gap in the electron density of BpaA affecting this region (Edwards *et al.*, 2010), thus presenting the first fully representative structure of this motif.

Owing to the considerable size of BPSL2063 (1090 residues) and the presence of just one conserved domain ($BPSL2063_{Xtal}$), the construction of additional domains mapping the entire polypeptide proved to be difficult (Edwards *et al.*, 2010). To facilitate this search, we constructed a single-gene domain-filtering library that allowed us to identify two main domains ($BPSL2063_{D1}$ and $BPSL2063_{D2}$) located at the N- and C-termini, respectively, which were produced as recombinant His-tagged proteins in soluble form for purification/crystallization trials. Notably, the majority of $BPSL2063_{Xtal}$ is covered by $BPSL2063_{D2}$.

The availability of a soluble/stable protein construct is often the rate-limiting factor in macromolecular crystallography and in protein-based research in general. It is estimated that only 30–40% of full-length proteins can be successfully expressed and purified; this figure decreases to ∼20% for soluble eukaryotic proteins (DiDonato *et al.*, 2004; Gräslund *et al.*, 2008; Phizicky *et al.*, 2003). In the case of BPSL2063, further to the identification of domains within one protein through a single-gene domain-filtering library, we aimed to develop a high-throughput tool to identify soluble protein domains from the entire protein repertoire of *B. pseudomallei* by generating a genomic DNA ORF-filtering library for analysis by NGS. By randomly fragmenting genomic DNA into 200–800 bp fragments and fusing them with the folding reporter TEM-1 β-lactamase, fragments that encode correctly folded/soluble proteins/domains can be specifically selected (D'Angelo *et al.*, 2011; Di Niro *et al.*, 2010; Zacchi *et al.*, 2003). NGS mapping reads obtained after genomic DNA library sequencing, and data regarding fragments that encode for ORF regions, provide unbiased information on the location of proteins/domains that are likely to result in soluble products. Our results imply that for complex bacterial genomes such as that of *B. pseudomallei* (two chromosomes, 7.2 Mb and 68% GC content), the ORF-filtering vector favours the selection of ORFs encoding for proteins that are potentially exposed on the cell surface, such as those involved in transport mechanisms or located in the cell membrane. Thus, the extension of the filtering power of our selection strategy to identify outer membrane-localized protein domains may be useful in the context of host–pathogen interaction studies and may support the identification of antigens. Soluble constructs were produced for three potential antigens represented in the ORF-filtered library. The structure of one such antigen (BPSL2520) has been recently solved by our group (to be published). Although we do not suggest that our approach will identify domains that will necessarily yield viable crystals, as there are numerous other biophysical factors that affect crystallogenesis, the identification of soluble/stable domains will undoubtedly increase the probability of successful crystal growth.

The advantage of our strategy that couples random genomic DNA fragmentation to filtering includes the fact that it does not require extensive analyses, gene-specific primer synthesis or PCR amplifications. Furthermore, the collection created consists of many different versions of each domain, each differing by a few amino acids. Identified protein fragments may find different applications, including functional studies, structural studies, antibody generation, protein–substrate binding analyses and domain shuffling for enzyme evolution. The filtering approach applied here to the *B. pseudomallei* genome could be applied to any other pathogen, leading to more direct cloning, expression and purification of soluble proteins/domains for any of the above applications, including crystallogenesis. The filtering strategy can also integrate well

electronic reprint

with *in silico*-based reverse vaccinology (RV) approaches to antigen discovery that rely heavily on successive antigen production for immunological testing.

Once a library has been generated, it can be exploited as a 'universal reagent', usually after the fragments have been recloned into a phage-display context. Once displayed on phage, fragments can be selected by using different targets. For example, by using sera from pathogen-infected patients the isolated fragments may allow the identification of bacterial antigens/epitopes specifically recognized by host antibodies. Conversely, interactome networks could be profiled by isolating fragments encoding domains carrying specific binding properties (*e.g.* the recognition of other proteins or domains; Di Niro *et al.*, 2010). Finally, when adequate probes are available (Evans & Cravatt, 2006; Velappan *et al.*, 2010), the identification of domains showing specific enzymatic activity could provide a simpler method to annotate genes (and their assignment to specific metabolic pathways) on the basis of determined function, rather than homology.

Acknowledgements

This project was supported by the Fondazione CARIPLO (Progetto Vaccini, contract No. 2009-3577), joint funding between the Fondazione CARIPLO and the Regione Lombardia (Progetto PROVA, contract No. 42666248). LJG was a recipient of Assegno di Ricerca (2012) from the University of Milan.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Boria, I., Boatti, L., Pesole, G. & Mignone, F. (2013). *BMC Bioinformatics*, **14**, Suppl. 7, S10.

Campos, C. G., Byrd, M. S. & Cotter, P. A. (2013). *Infect. Immun.* **81**, 2788–2799.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Cheng, A. C. & Currie, B. J. (2005). *Clin. Microbiol. Rev.* **18**, 383–416.

D'Angelo, S., Mignone, F., Deantonio, C., Di Niro, R., Bordoni, R., Marzari, R., De Bellis, G., Not, T., Ferrara, F., Bradbury, A., Santoro, C. & Sblattero, D. (2013). *Clin. Immunol.* **148**, 99–109.

D'Angelo, S., Velappan, N., Mignone, F., Santoro, C., Sblattero, D., Kiss, C. & Bradbury, A. R. (2011). *BMC Genomics*, **12**, Suppl. 1, S5.

Dautin, N. & Bernstein, H. D. (2007). *Annu. Rev. Microbiol.* **61**, 89–112.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, J. S. (2007). *Nucleic Acids Res.* **35**, W375–W383.

Desvaux, M., Cooper, L. M., Filenko, N. A., Scott-Tucker, A., Turner, S. M., Cole, J. A. & Henderson, I. R. (2006). *FEMS Microbiol. Lett.* **264**, 22–30.

DiDonato, M., Deacon, A. M., Klock, H. E., McMullan, D. & Lesley, S. A. (2004). *J. Struct. Funct. Genomics*, **5**, 133–146.

Di Niro, R., Sulic, A. M., Mignone, F., D'Angelo, S., Bordoni, R., Iacono, M., Marzari, R., Gaiotto, T., Lavric, M., Bradbury, A. R., Biancone, L., Zevin-Sonkin, D., De Bellis, G., Santoro, C. & Sblattero, D. (2010). *Nucleic Acids Res.* **38**, e110.

Edwards, T. E., Phan, I., Abendroth, J., Dieterich, S. H., Masoudi, A., Guo, W., Hewitt, S. N., Kelley, A., Leibly, D., Brittnacher, M. J., Staker, B. L., Miller, S. I., Van Voorhis, W. C., Myler, P. J. & Stewart, L. J. (2010). *PLoS One*, **5**, e12803.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Evans, M. J. & Cravatt, B. F. (2006). *Chem. Rev.* **106**, 3279–3301.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Felgner, P. L. *et al.* (2009). *Proc. Natl Acad. Sci. USA*, **106**, 13499–13504.

Gaudesi, D., Peri, C., Quilici, G., Gori, A., Ferrer-Navarro, M., Conchillo-Sole, O., Thomas, R., Nithichanon, A., Lertmemongkolchai, G., Titball, R., Daura, X., Colombo, G. & Musco, G. (2015). *ACS Chem. Biol.* **10**, 803–812.

Giuliani, M. M. *et al.* (2006). *Proc. Natl Acad. Sci. USA*, **103**, 10834–10839.

Gourlay, L. J. *et al.* (2013). *Chem. Biol.* **20**, 1147–1156.

Gourlay, L. J., Thomas, R. J., Peri, C., Conchillo-Solé, O., Ferrer-Navarro, M., Nithichanon, A., Vila, J., Daura, X., Lertmemongkolchai, G., Titball, R., Colombo, G. & Bolognesi, M. (2015). *FEBS J.* **282**, 1980–1997.

Gräslund, S. *et al.* (2008). *Nature Methods*, **5**, 135–146.

Hartmann, M. D., Grin, I., Dunin-Horkawicz, S., Deiss, S., Linke, D., Lupas, A. N. & Hernandez Alvarez, B. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 20907–20912.

Heger, A. & Holm, L. (2003). *J. Mol. Biol.* **328**, 749–767.

Henderson, I. R., Navarro-Garcia, F. & Nataro, J. P. (1998). *Trends Microbiol.* **6**, 370–378.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 133–144.

Lassaux, P., Peri, C., Ferrer-Navarro, M., Gourlay, L. J., Gori, A., Conchillo-Solé, O., Rinchai, D., Lertmemongkolchai, G., Longhi, R., Daura, X., Colombo, G. & Bolognesi, M. (2013). *Structure*, **21**, 167–175.

Lazar Adler, N. R., Stevens, J. M., Stevens, M. P. & Galyov, E. E. (2011). *Front. Microbiol.* **2**, 151.

Malito, E. & Rappuoli, R. (2013). *Chem. Biol.* **20**, 1205–1206.

Meng, G., St Geme, J. W. III & Waksman, G. (2008). *J. Mol. Biol.* **384**, 824–836.

Meng, G., Surana, N. K., St Geme, J. W. & Waksman, G. (2006). *EMBO J.* **25**, 2297–2304.

Nummelin, H., Merckel, M. C., Leo, J. C., Lankinen, H., Skurnik, M. & Goldman, A. (2004). *EMBO J.* **23**, 701–711.

Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M. & Fields, S. (2003). *Nature (London)*, **422**, 208–215.

Quinlan, A. R. & Hall, I. M. (2010). *Bioinformatics*, **26**, 841–842.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna: The R Foundation for Statistical Computing.

Suwannasaen, D., Mahawantung, J., Chaowagul, W., Limmathurotsakul, D., Felgner, P. L., Davies, H., Bancroft, G. J., Titball, R. W. & Lertmemongkolchai, G. (2011). *J. Infect. Dis.* **203**, 1002–1011.

Szczesny, P., Linke, D., Ursinus, A., Bär, K., Schwarz, H., Riess, T. M., Kempf, V. A., Lupas, A. N., Martin, J. & Zeth, K. (2008). *PLoS Pathog.* **4**, e1000119.

Szczesny, P. & Lupas, A. (2008). *Bioinformatics*, **24**, 1251–1256.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Velappan, N., Fisher, H. E., Pesavento, E., Chasteen, L., D'Angelo, S., Kiss, C., Longmire, M., Pavlik, P. & Bradbury, A. R. (2010). *Nucleic Acids Res.* **38**, e22.

Zacchi, P., Sblattero, D., Florian, F., Marzari, R. & Bradbury, A. R. (2003). *Genome Res.* **13**, 980–990.

electronic reprint