Published online 19 July 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 18* ***e125***
doi:10.1093/nar/gkr548

Experimental mapping of soluble protein domains using a hierarchical approach

Jean-Denis Pedelacq[1,2,*], Hau B. Nguyen[3], Stephanie Cabantous[4,5,6], Brian L. Mark[7], Pawel Listwan[3], Carolyn Bell[3], Natasha Friedland[3], Meghan Lockard[3], Alexandre Faille[1,2], Lionel Mourey[1,2], Thomas C. Terwilliger[3] and Geoffrey S. Waldo[3,*]

[1]CNRS; IPBS (Institut de Pharmacologie et de Biologie Structurale), 205 route de Narbonne, F-31077 Toulouse, France, [2]Université de Toulouse; UPS, IPBS, F-31077 Toulouse, France; [3]Bioscience Division, MS-M888, Los Alamos National Laboratory, Bikini Atoll Rd, SM30, Los Alamos, NM 87545, [4]INSERM UMR1037-Cancer Research Center of Toulouse, [5]Université de Toulouse, [6]Institut Claudius Régaud, 31052 Toulouse Cedex and [7]Department of Microbiology, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

Received April 12, 2011; Revised May 26, 2011; Accepted June 15, 2011

ABSTRACT

Exploring the function and 3D space of large multidomain protein targets often requires sophisticated experimentation to obtain the targets in a form suitable for structure determination. Screening methods capable of selecting well-expressed, soluble fragments from DNA libraries exist, but require the use of automation to maximize chances of picking a few good candidates. Here, we describe the use of an insertion dihydrofolate reductase (DHFR) vector to select in-frame fragments and a split-GFP assay technology to filter-out constructs that express insoluble protein fragments. With the incorporation of an IPCR step to create high density, focused sublibraries of fragments, this cost-effective method can be performed manually with no *a priori* knowledge of domain boundaries while permitting single amino acid resolution boundary mapping. We used it on the well-characterized p85α subunit of the phosphoinositide-3-kinase to demonstrate the robustness and efficiency of our methodology. We then successfully tested it onto the polyketide synthase PpsC from *Mycobacterium tuberculosis*, a potential drug target involved in the biosynthesis of complex lipids in the cell envelope. X-ray quality crystals from the acyl-transferase (AT), dehydratase (DH) and enoyl-reductase (ER) domains have been obtained.

INTRODUCTION

Over the past 10 years, the *Mycobacterium tuberculosis* Structural Genomics Consortium (http://www.doe-mbi.ucla.edu/TB/), a large-scale center funded by the National Institutes of General Medical Sciences (NIGMS), has cloned more than 1400 protein targets for cell-based production in *Escherichia coli*. Only half of the proteins expressed have been produced in a soluble form. Other structural genomics initiatives also confirmed this step to be one major bottleneck in structural biology. Screening approaches for improved folding and stability of protein targets have brought new insights into solving structures of single-domain proteins (1,2). However, structure determination of multidomain proteins has been found to be more challenging due to their larger size and increased instability. In addition, domain boundaries are not always straightforward to predict (3). In this respect, high-throughput approaches of generating libraries of truncated DNA fragments (4) combined with a colony filtration immunoblot using antibody detection of tagged constructs (5,6), the detection of a fluorescent fused GFP phenotype (7,8) or a fused C-terminal biotin acceptor peptide (9) have proven to be successful in identifying constructs potentially amenable to functional and structural characterization. Unfortunately, all these approaches lack a filtering strategy to effectively eliminate DNA fragments, which do not encode authentic protein domains. Instead, fully automated strategies have been implemented to effectively screen the thousands of clones and pick a few good candidates (10).

*To whom correspondence should be addressed. Tel: +33 5 61 17 54 11; Fax: +33 5 61 17 59 94; Email: Jean-Denis.Pedelacq@ipbs.fr
Correspondence may also be addressed to Geoffrey S. Waldo. Tel: +505 667 8161; Fax: +505 665 3024; Email: waldo@lanl.gov

Eliminating frame-shifted fragments from DNA libraries has become a key step to addressing the construction of expression plasmid libraries that produce protein domains in-frame with the target gene. Existing systems involve expressing fragments as N-terminal fusions to murine dihydrofolate reductase (mDHFR) (11), kanamycin (12) or β-lactamase (13). However these technologies yield false positives originating from translation initiation at internal ribosome binding sites (IRBS). A number of bipartite selection systems have also been developed in an attempt to overcome these limitations. In these systems, the DNA sequence of interest is inserted between the two halves of the reporter, which are both required to give an observable phenotype (14–18).

In this article, we describe a novel approach that uses a two-body *E. coli* dihydrofolate reductase (DHFR) (19) scaffold for selecting in-frame DNA sequences from a random library of a fragmented gene, combined with the split-GFP technology (20) to identify soluble candidates. We used the regulatory subunit p85α of the class I_A phosphoinositide 3-kinase (PI3K) as a benchmark to validate our method. We then tested it onto the polyketide synthase PpsC from *M. tuberculosis*. This 230 kDa mega-synthase plays a key role in the virulence of this microbial pathogen through the synthesis of phtiocerol dimycocerosates, a family of lipids located in the cell envelope. With the incorporation of an inverse polymerase chain reaction (PCR) step to increase population density in fragments within domains, X-ray quality crystals from the acyl-transferase (AT), dehydratase (DH) and enoyl-reductase (ER) domains have been obtained.

MATERIALS AND METHODS

Gene cloning and fragmentation

The *p85α* and *Ppsc* genes from *M. tuberculosis* were cloned into the NdeI/BamHI and NdeI/SpeI sites of a pET26b plasmid (Novagen, Madison, WI, USA), respectively, and PCR amplified using gene-specific primers (Supplementary Data 2). DNA fragmentation conditions of the *p85α* gene were optimized using small aliquots of concentrated PCR products incubated with a serial 2-fold dilutions of a DNase I stock solution at 1 U/µl (Invitrogen, Carlsbad, CA, USA). Best condition corresponded to a 24-fold dilution from a 1 µl stock solution of DNase I with 20 µl of 10 mM Tris–Hcl pH = 7.4 and 3 µl of 10 mg/ml Bovine Serine Albumine (BSA). Cleaned PCR product of 50 µl was mixed with 6 µl of 0.5 M Tris–Hcl pH = 7.4 and 1 µl of 100 mM $CoCl_2$. Both solutions were pre-incubated in a PCR block at 15°C for 5 min before mixing. Two libraries were generated by adding 6 µl of the 24-fold DNase I solution to the PCR mixture: a 250–400 bp DNA library (small size) and a 400–750 bp DNA library (large size) with incubation times of 5 and 2 min, respectively. Digestion reactions were stopped by adding 650 µl of PB buffer before cleaning through a Qiaquick PCR purification column (Qiagen Inc. USA, Valencia, CA, USA). In the case of *Ppsc*, small size (400–850 kb) and large size (850–1650 kb) DNA libraries were obtained from 160 µl of cleaned PCR product using a HydroShear device from Genomics Solutions (Ann Arbor, MI, USA) applying 25 cycles at speed codes 8 and 13, respectively. Extremities of the fragments were polished using 3′-5′-exonuclease activity of Vent polymerase (New England Biolabs, Beverly, MA, USA) at 72°C for 20 min. Double-stranded DNAs were resolved by agarose gel electrophoresis and visualized by ethidium bromide staining. A slab of gel containing DNA fragments with desired size was then excised and recovered with a QIAquick gel extraction kit (Qiagen Inc. USA, Valencia, CA, USA). DNA fragments designed for the screening of BCR domain constructs were amplified using gene-specific primers (Supplementary Data 2) and ligated into the NdeI/BamHI of pTET ColE1 GFP 11 vector.

Construction of insertion DHFR library

Blunt fragments were ligated in a StuI-digested insertion DHFR (iDHFR) pET vector (Supplementary Data 1) for 12 h at 16°C, and ligated plasmids were transformed into electro-competent *E. coli* DH10B cells (Invitrogen, Carlsbad, CA, USA) to increase efficiency. Starting with 5 × 106 clones/library, transformed cells were plated onto Luria-Bertani (LB) agar plates containing 35 µg/ml kanamycin, allowing the *E. coli* cells lawn to grow overnight at 37°C. Overnight colonies from lawns of 3 × 105 clones, estimated by dilution plates, were washed off and used for plasmid preparation prior to transformation into chemically competent *E. coli* BL21 (DE3) Tuner cells. Following overnight growth at 32°C on LB/kanamycin medium, cells were diluted in LB containing 20% glycerol to $OD_{600\,nm} = 1.0$ for −80°C freezer stocks. Forty microliter of the 1.0 OD freezer stock were used to seed a 3 ml LB/kanamycin culture. Cells were propagated until $OD_{600\,nm} = 0.5$ was reached, and then induced with 20 µM IPTG for an additional 2–3 h. Cells were diluted in 1 ml LB to $OD_{600\,nm} = 2.0$, yielding 10 × 108 cells/ml, and plated on LB medium containing 6 µg/ml trimethoprim (TMP) and 20 µM IPTG. To compare the colony-forming unit (CFU), numbers in the presence or absence of TMP, cells were further diluted to 1/16000 and plated out on two LB/agar plates containing 20 µM IPTG in the presence or absence of TMP. All plates were incubated overnight at 32°C.

Inverse PCRs

Recovered iDHFR libraries were diluted for plasmid preparation. NdeI/BamHI and NdeI/SpeI restrictions sites were used to release fragments from *p85α* and *Ppsc*, respectively. Gel extracted and cleaned inserts were ligated into their corresponding digested pTET ColE1 GFP 11 vector. Inverse PCRs were performed following the protocol by Hoskins and colleagues (21). For each *p85α* and *Ppsc* targeted domain, phosphorylated forward and reverse primers were designed (Supplementary Data 2). Briefly, 100 µl inverse PCRs (IPCRs) were conducted with Phusion DNA polymerase (Finnzymes) according to the manufacturer's instructions. Following a self-ligation with T4 DNA ligase in a 100 µl volume at 16°C overnight and a digestion with DpnI enzyme at 37°C for 2 h 30 min, ligated pTET ColE1 GFP 11 vectors containing targeted inserts were transformed into chemically competent BL21 (DE3) pET GFP 1–10 cells.

Solubility screens using the split-GFP assay

In vivo solubility screenings were performed as previously described (22). Briefly, cells were grown to saturation in LB containing 35 µg/ml kanamycin and 75 µg/ml spectinomycin, and diluted in 20% glycerol to $OD_{600\,nm} = 1.0$ for −80°C freezer stocks. Frozen cells were thawed at 0°C, 400-fold diluted (twice) in LB and plated onto a nitrocellulose membrane with selective LB-agar containing the same antibiotics (approximately 3000 colonies). After overnight growth at 32°C, the membrane was transferred onto a pre-warmed plate containing 250 ng/ml AnTet for 2 h, and rested back onto its original LB-Kan-Spec plate for 1 h. Following induction with 1 mM IPTG at 37°C for 1 h, the induced colonies were illuminated using an Illumatool Lighting System (LightTools Research), equipped with a 488 nm excitation filter. An ensemble of 96 clones with decreasing levels of *in vivo* fluorescence intensities were picked for each library. Columns 1–3 of the tissue culture plate correspond to only bright clones, columns 4–9 present a range in fluorescence intensity levels from medium bright to faint and columns 10–12 only correspond to very faint clones. As a control, a total of 96 clones were picked randomly by hand, transferred to 96-well plates and grown before sequencing. All the clones were used as starter cultures on 96-well tissue culture plates for *in vitro* complementation split-GFP assays using our in-house automated, high-throughput, liquid-handling platform (23).

Identification of fragments boundaries

Individually picked clones were grown overnight at 30°C in a 96-well tissue culture plate containing 7.5% glycerol in LB-Kan-Spec medium. Plasmid amplification at the Los Alamos genome sequencing facility using rolling-circle amplification in the presence of a forward primer specific of *tet*-promoter and a reverse primer specific of GFP 11 (Supplementary Data 2) yield high-quality sequence. DNA sequences were analyzed using BioEdit® software. Sequence alignments were performed by aligning individual fragments onto the full-length parent gene to determine the exact boundaries from the forward (start of the fragment) and reverse sequence (end of the fragment). Based on the *in vitro* solubility assays, fragments were color-coded black, light green and bright green, where the black side of the spectrum identifies the least soluble protein fragments and the bright green side corresponds to the top 25% of the most soluble ones. Fraction of color-coded black and light green fragments varies from 25% and 50% for a full-length gene mapping to 37.5% in the case of IPCRs.

Small scale expression and solubility tests

An ensemble of 10 fragments spanning the PpsC polypeptide chain were selected and subcloned from the pTET-GFP, 11 plasmid into a N6–HIS or C6–HIS pET vector. The resulting clones were grown at 37°C in 1 ml cultures using 35 µg/ml kananycin. Cells were induced in exponential phase with 1 mM IPTG for 3 h. Cell culture pellets of 1 ml of each fragment were separately resuspended in 40 µl 150 mM NaCl, 100 mM Tris–HCl pH = 7.5, 10% (v/v) glycerol (TNG buffer) and sonicated. The lysate was fractionated by centrifugation to yield the soluble and pellet fractions. The pellet fraction was washed twice with 100 µl TNG, centrifuged and resuspended in the same starting volume. Samples corresponding to the soluble (S) and pellet (P) fractions were resolved on a 4–20% gradient Criterion SDS–PAGE gel (Bio-Rad, Hercules, CA, USA). Protein samples were stained using Gel Code Blue stain reagent (Pierce, Rockford, IL) and imaged using a GS-800 Calibrated Densitometer (Biorad, Hercules, CA, USA).

Metal affinity resin purification of selected fragments

Five hundred milliliter cultures of BL21(DE3) cells expressing selected protein fragments were grown to OD (600 nm) ∼0.5–0.7 in LB medium supplemented with 1 mM kanamycin, induced with 0.5 mM IPTG for 5 h at 32°C, pelleted by centrifugation, resuspended in 15 ml 100 mM Tris–Hcl pH = 8.1 containing 150 mM NaCl and sonicated. The soluble extract of 15 ml was mixed with an equal volume of 50% v/v slurry of metal affinity resin beads (Talon resin, Clontech, Palo Alto, CA, USA) in TNG buffer for 10 min and centrifuged briefly. The unbound fraction was removed by pipetting and the beads were washed twice with 10 volumes of TNG loading buffer. After an additional wash with TNG buffer supplemented with 10 mM imidazole, His-tagged proteins were eluted with 250 mM imidazole in TNG buffer. For each purification step, the proteins elution samples were resolved on a 4–20% gradient Criterion SDS–PAGE gel (Bio-Rad, Hercules, CA, USA) and stained using the same procedure described above. Using this procedure, 98% pure AT (∼4 mg), DH (∼20 mg) and ER (∼5 mg) proteins were obtained. The absence of aggregates in the samples and polydispersity levels of <10% were confirmed using a DynaPro^{TM} Dynamic Light Scattering (DLS) Instrument from Wyatt Technology.

^{15}N–Protein labeling and sample preparation for NMR spectroscopy

Protein expression level in minimal media was enhanced by increasing cell density using a 4:1 cell concentrating method (24). For ^{15}N uniform labeling, cells from 21 of LB media were grown at 37°C until an OD_{600} of 0.5–0.7 was reached, then harvested and resuspended in 500 ml of M9 minimal media containing 0.5 g $^{15}NH_4Cl$. Following an additional 30 min of shaking, cells were induced with 0.5 mM IPTG for 7 h at 25°C. Proteins were purified as described above using Talon resin and dialyzed against 50 mM Na phosphate buffer pH = 7, 1 mM DTT, 1 mM EDTA overnight at 4°C to remove imidazole and salt from elution buffer. Protein solution was concentrated using an Amicon Ultra-15 centrifugal filter device (10 kDa cutoff; Millipore). Final NMR samples usually contain 0.5–1 mM protein in 50 mM Na phosphate buffer pH = 7, 1 mM DTT, 1 mM EDTA and 10% D2O. ^{1}H–^{15}N HSQC spectra were recorded at 298 K on

Varian Inova 720 MHz spectrometer using a conventional probe. Complex points of 1024 in the direct dimension (^{1}H) and 256 complex points in the indirect dimension (^{15}N) were collected. All spectra were processed using nmrPipe (25) and analyzed by the Sparky software (Goddard and Kneller, University of California, San Francisco). Chemical shifts were referenced to 4,4-dimethyl-4-silapentane-1-sulfonic acid (26).

RESULTS

Selection of in-frame, well-expressed and soluble fragments

Our strategy has four distinct steps, where selection pressure forces the false positives at any given step to be effectively eliminated and the number of false negatives to be reduced. First, a library of 5 × 106 clones expressing DNA fragments is created by fragmentation of a PCR amplified gene using DNase I or mechanical shearing (Figure 1). Fragments of the desired size are excised from preparative agarose gel, blunt-ended with 5′-3′-exonuclease, and then cloned between the two halves of bacterial DHFR at a permissive site between amino acids 86 and 87 in the presence of TMP (Supplementary Data 1). We found that 3–6 µg/ml TMP killed DH10B *E. coli* cells, but allowed clones expressing inserts without stop codons to subsist. Under this selective antibiotic pressure, approximately 1 in 18 cells survive, corresponding to fragments translated in the same reading frame as the parent open reading frame (ORF) and in-frame with the reporter destination vector (8). This leads to 3 × 105 clones recovered at this step (Figure 1), a number to compare with the few hundreds in-frame clones from published methods (5,6,11). Only fully automated platforms capable of picking and assaying tens of thousands of clones in parallel for expression and solubility can compete with our approach. In this respect, Expression of Soluble Proteins by Random Incremental Truncation (ESPRIT) (9) has led to remarkable results on several challenging targets (10,27–30). In the third step, in-frame fragments are subcloned into the split-GFP system. At this point, full-length libraries can be screened for solubility after *in vivo* sequential induction of the GFP 11-tagged protein fragments and the complementary GFP 1–10 detector (22). Clones displaying a wide

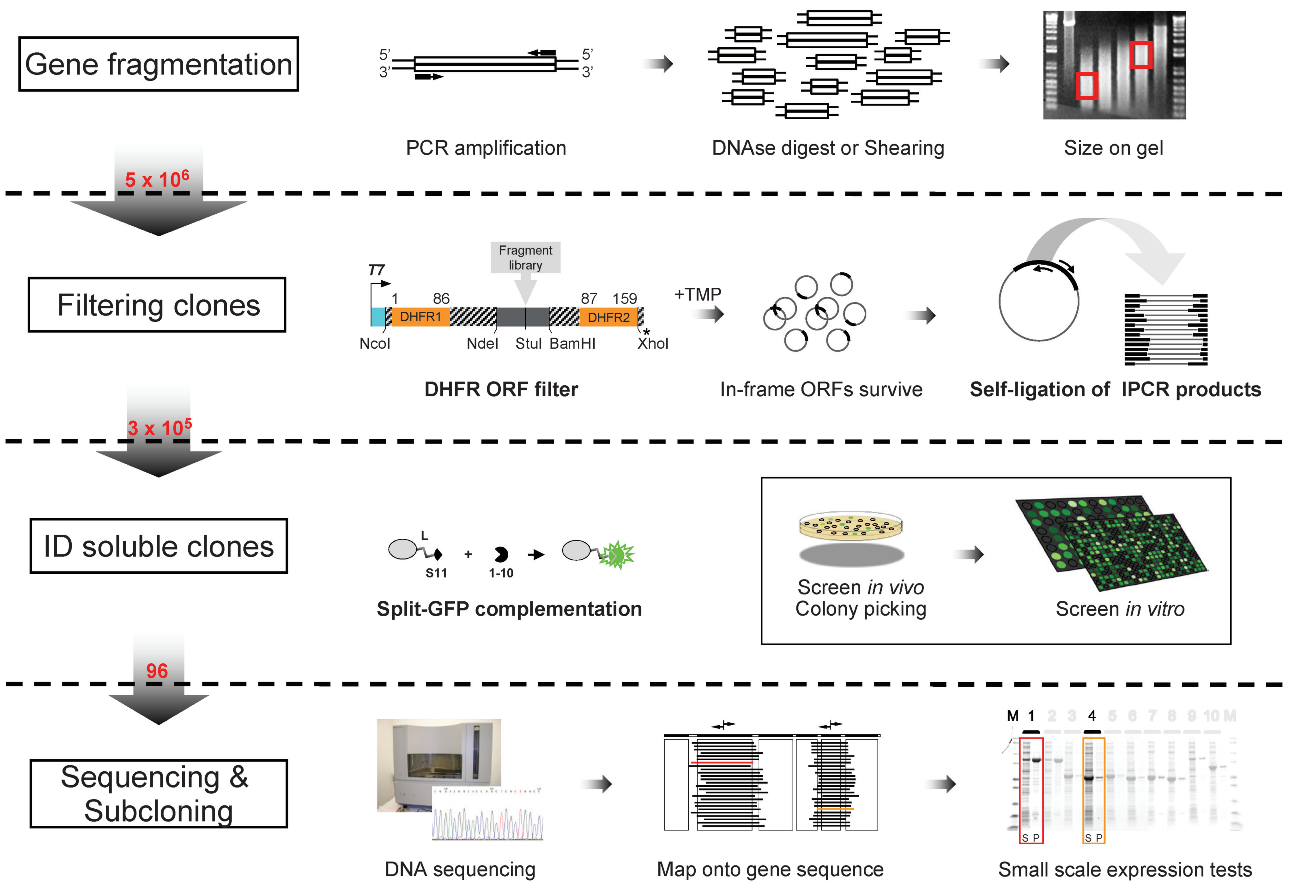

Figure 1. The GFP-enabled domain trapping strategy. The PCR-amplified gene is fragmented by chemical or mechanical means and DNA fragments of desired size are excised from agarose gel. Blunt-end fragments are cloned into the iDHFR ORF filter, where only the in-frame ones permitting the expression of the second half of DHFR will survive. Inserts from the recovered plasmids are cloned into the split-GFP vector and used for IPCR to create high density, focused sublibraries of fragments prior to the split-GFP assay. A range of fluorescent clones are picked and grown in 96-well liquid cultures for *in vitro* quantification of the soluble and insoluble protein fractions. Clones are sequenced and the fragments are aligned onto the full parent gene. Fragments can be directly tested for expression or subcloned without the S11 tag into a pET vector. Numbers to the left indicate the approximate library size at the different steps.

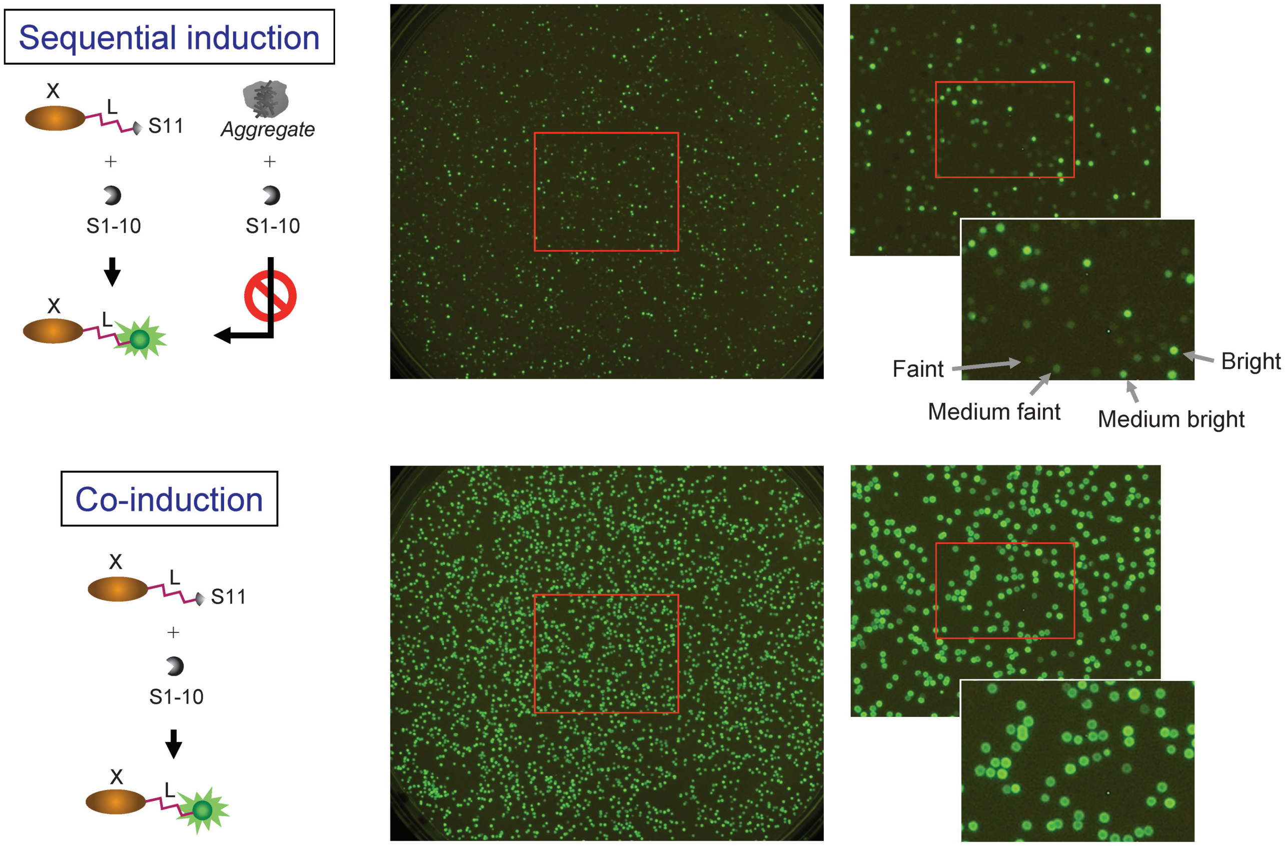

Figure 2. Screening clones using the split-GFP reassembly assay. Images showing cell colony fluorescence from agar plates after sequential induction (solubility reporter) and co-induction (expression reporter) of the GFP 11-tagged protein fragments and its complementary GFP 1–10 detector. Clones displaying a wide range of fluorescence are visible after sequential induction.

range of fluorescence are picked from the agar plates and grown in 96-well liquid culture plates (Figure 2). Within each library, we picked a total of 384 clones in four 96-well plates with fluorescence intensity levels from bright to faint (see Methods in Supplementary Data). As an alternative, IPCR (21) can first be used to create high density, focused sublibraries of fragments prior to the split-GFP assay. In this case, 96 clones were picked per sublibrary of fragments to ensure maximum coverage (Figure 1). For *in vitro* quantification, *E. coli* cells were first induced with anhydrotetracycline (AnTET) to overexpress the GFP 11-tagged proteins. Soluble lysates and insoluble fractions were assayed by adding the GFP 1–10 detector fragment, as previously described (22). In the final step, DNA sequencing was used to determine the boundaries of each fragment by reference to the parent gene. *In silico*, all fragment sequences within a library were aligned onto the parent gene and color-coded by solubility levels. We used the color scheme *black/light green/bright green*, where the black side of the spectrum identifies the least soluble fragments and the bright-green color identifies the most soluble ones. This provides a visual tool to correlate fragment boundaries with solubility levels and makes it easy to identify the most compact and soluble fragments for downstream functional and structural characterization.

p85α as a benchmark for testing our domain trapping strategy

The structural organization of the regulatory subunit p85α of the class I_A PI3K has been very well studied. Except for the two coiled-coil regions CC1 and CC2, the documented structures of the well-folded SH3 (31), BCR (32), N–SH2 (33) and C–SH2 (34) make p85α a good benchmark for demonstrating the feasibility and efficiency of domain trapping strategies.

Large quantities of PCR amplified target DNA (1–2 µg) were produced using either the Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA, USA) or the Phusion® High-Fidelity DNA polymerase (NEB, Ipswich, MA, USA). Since DNA ligation efficiency varies inversely with the size of the insert, and to avoid biased ligation of small fragments, two individual libraries of fragments were created: from 250 to 500 bp and from 350 to 750 bp. Indeed, the larger size fragment library would let us 'fish' soluble fragments encompassing the larger BCR domain (550 bp), whereas the smaller size fragment library would favor the selection of smaller soluble fragments from the SH3 domain (252 bp) and the two SH2 domains (345 and 324 bp for N–SH2 and C–SH2, respectively). DNase I reaction conditions were optimized to narrow the window of highly concentrated DNA fragments in the desired size range (see Methods in Supplementary Data).

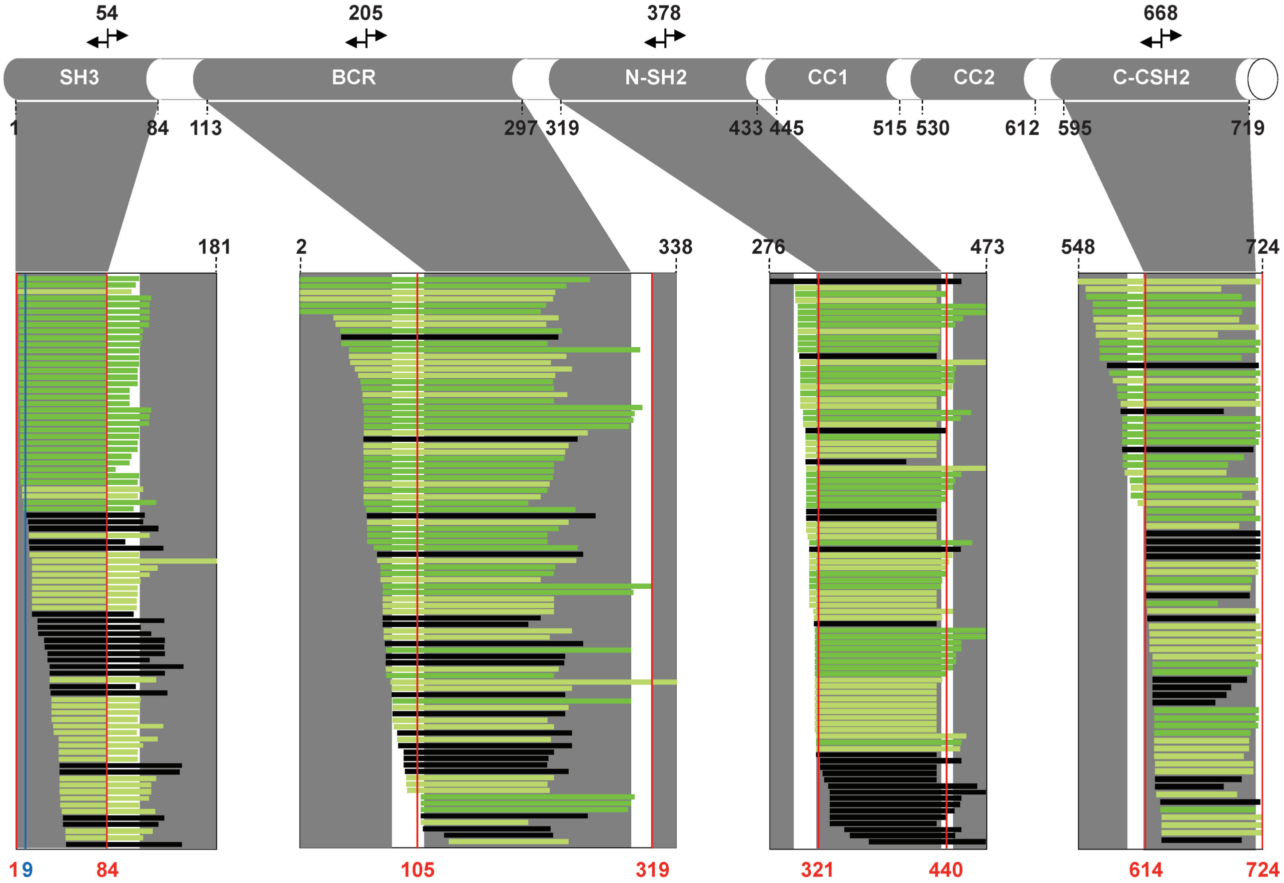

Figure 3. Mapping of the IPCR p85α targeted fragments. IPCRs using oppositely directed primers were used to generate sublibraries of fragments. For each sublibrary, an ensemble of 96 clones with a wide range of *in vivo* fluorescence intensities were picked and grown in 96-well liquid culture plates for *in vitro* split-GFP solubility screen. Only the correctly sequenced in-frame fragments are represented. Based on the *in vitro* solubility assays, fragments were color-coded black, light green and bright green, where the black side of the spectrum identifies the bottom 20% least soluble protein fragments and the bright green side corresponds to the top 20% most soluble ones. IPCRs within p85α were used to generate four large-size sublibraries (350–750 kb) centered onto the SH3, BCR, N–SH2 and C–SH2 domains. Within each library, solubility values from three or more identical fragments were averaged in order to keep the color-coded representation as clear as possible. Boundaries of structure solved domains are indicated in red and the junction at amino acid position 9 is indicated in blue.

The two independent pools of blunted *p85α* fragments were cloned into the iDHFR vector and transformed into *E. coli* DH10B electro-competent cells (Invitrogen, Carlsbad, CA, USA). Transformation of the recovered plasmids into *E. coli* BL21 Tuner ^{TM}(DE3) competent cells (Novagen, San Diego, CA, USA) containing the *lacY* permease mutation allowed the uniform induction of the cells by IPTG during the subsequent selection of in-frame clones. To ensure that the selection process is consistent with the theoretical 1 in 18 clones expected to survive after the ORF-filter step, each library was diluted and plated on medium in the presence and absence of TMP for fast and accurate colony counting.

Initially, pools of 'in-frame' DNA fragments were subcloned into the pTET-GFP 11 solubility vector using NdeI and BamHI sites (Supplementary Data 1), and transformed into BL21 (DE3) cells containing the pET GFP 1–10 plasmid to screen for soluble expression (20). After sequential induction of the GFP 11-tagged fragments, we could take advantage of the split-GFP complementation assay to screen thousands of clones *in vivo* by visual assessment of their intrinsic fluorescence, which is well correlated with the amount of soluble protein expressed (Figure 2). Both fragment libraries displayed substantial phenotypic variability as observed from the wide distribution of fluorescence intensities. Even though sequencing information was missing for some of the 384 manually picked clones (79 for the small size and 16 for the large size library), probably due to cross contamination with nearby faint or black clones, only a limited number of the sequenced fragments were not in the authentic reading frame (9 out of 305 for the small size and 2 out of 368 for the large size library), thus illustrating the efficiency of selection process. Despite the fact that most p85α domains are well represented, with the exception of the BCR domain, we noticed a bias in the distribution of fragments towards the second-half of the gene, from the N–SH2 domain to the C-terminal end of the gene (Supplementary Figure S1). In contrast, the distribution of randomly picked fragments is more homogeneous, as no information concerning fluorescence intensity levels was taken into consideration (data not shown).

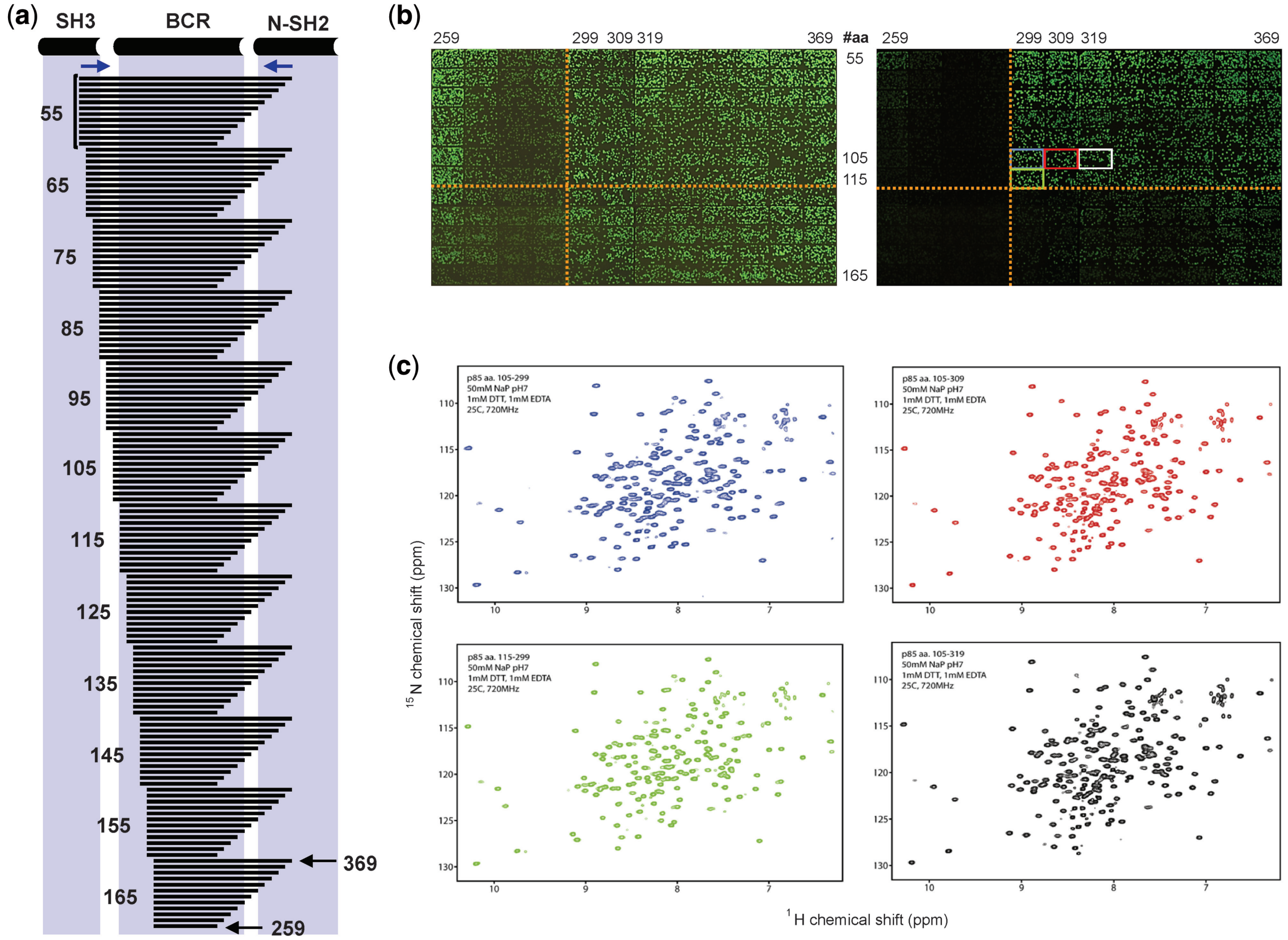

Figure 4. PCR-directed truncations of the p85α BCR domain. (**a**) Schematic representation of all 144 constructs aligned onto the p85α amino-acid sequence. N- and C-terminal positions are indicated. Fragments are organized in groups of 12 with identical N-terminal positions. (**b**) Expression (left) and solubility (right) levels of *E. coli* BL21(DE3) cells expressing the corresponding fragments in fusion with S11 following complementation with GFP 1–10. Orange dashed lines mark boundaries, where dramatic changes in expression and solubility levels were observed. (**c**) In addition to 105–319 for which the X-ray structure is known (PDB code: 1PBW), well-expressed and soluble fragments were selected for downstream NMR studies. HSQC spectra of fragments 105–299 (blue), 105–309 (red), 105–319 (black with corresponding white rectangle in the solubility screen) and 115–299 (green) are represented.

We reasoned that the bias may originate from our multistep selection process that favors the selection of extremely well-behaved fragments to the detriment of less soluble ones. To circumvent this representation artifact and starting with the same original library of in-frame fragments (following iDHFR selection), we used the IPCR (21) technique to selectively enrich for DNA sequences in regions of *p85α*. Phosphorylated primers were designed (Supplementary Data 2) to generate three small-size sublibraries (250–400 bp) centered on the SH3, N–SH2 and C–SH2 domains and one large-size sublibrary (400–750 bp) centered onto the BCR domain (Figure 3). We noticed that cutting into the N-terminal region of the SH3 and N–SH2 domains had a dramatic effect onto the solubility levels of the selected fragments. Also, the incorporation of amino acid residues from the CC2 domain and linker region drastically improved the solubility of the C–SH2 centered fragments, a somewhat surprising and difficult result to predict considering the unstructured nature of these regions.

Fragments can occasionally be used as a starting point to identify more compact and soluble constructs with a clear objective of maximizing the chances of 3D structure (35). As illustrated in Figure 3, IPCR sublibrary centered on the BCR domain contains three soluble fragments (110–294, 110–297 and 110–300) with C-terminal positions slightly shorter than the structurally characterized fragments 105–319 (PDB code: 1PBW). We note that the low solubility of another BCR fragment (117–297) from the full-length *p85α* library could be attributed to its shorter N-terminal end (Supplementary Figure S1b). In an effort to rationalize the effect of N- and C-terminal truncations in the BCR region, we cloned an ensemble of 144 constructs corresponding to a 10 amino acids walk from positions 55 and 369 (Figure 4 and Supplementary Data 2). Protein solubility levels were assessed both *in vivo* and *in vitro* using the split-GFP assay (20). Figure 4 illustrates the expression and solubility levels of all 144 constructs, respectively. As expected, fragments are soluble when the BCR core domain is

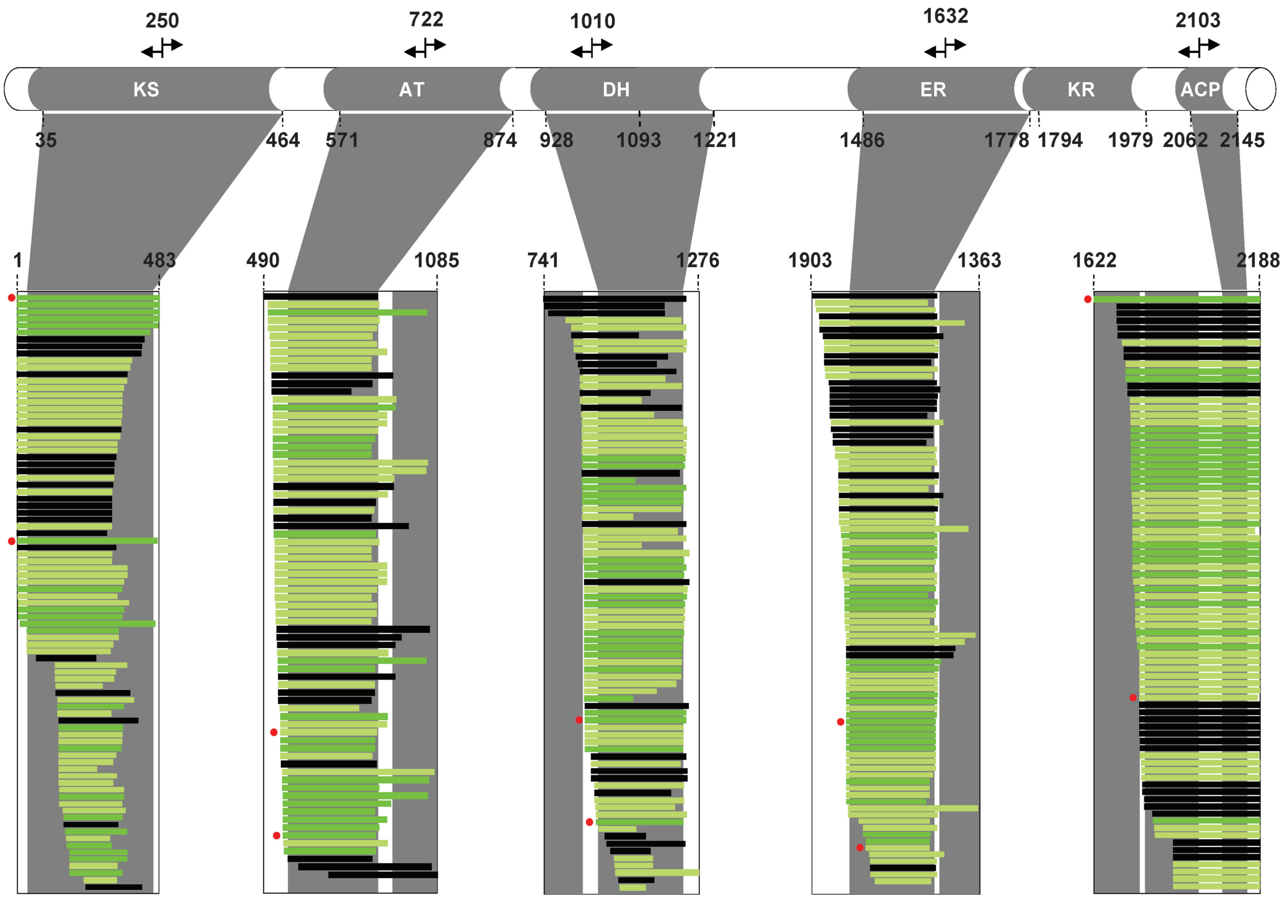

Figure 5. Mapping of the IPCR PpsC targeted fragments. Five large-size sublibraries of fragments (850–1650 kb) centered onto the KS, AT, DH, ER and ACP domains were generated. Picking and color coding of fragments follow the same rules as in Figure 3. Red dots indicate the fragments selected for downstream biochemical and structural characterization.

present. As truncations within the core proceed, protein solubility levels fall. Approximate boundaries of soluble fragments range between positions 115 and 125 in the N-terminal region, and positions 289 and 299 in the C-terminal region. A finer screening using one amino acid incremental truncation from the N-terminal (116–123) and C-terminal (290–297) regions was then performed (Supplementary Figure S2). Based on the fluorescence intensity ratios of solubility over expression, the fragments 116–294 is the shortest most soluble fragment of the BCR domain. Interestingly, cutting into the C-terminal end of helix $\alpha10$, from positions 297 to 294, did not seem to have an effect on the solubility of the last three fragments within the group. The addition of a three residue linker (Gly–Ser–Asp) to the C-terminus of the protein as a result of molecular cloning site in the pTET-GFP 11 (20) vector may compensate for residues Ser–Thr–Glu of the real protein sequence, thus making truncated variants more stable. A total of six most compact soluble protein fragments, including the structurally characterized fragments 105–319 (PDB code: 1PBW), were selected and purified for 2D NMR experiments. ^{1}H–^{15}N HSQC spectra are well resolved and dispersed, thus indicating that all protein fragments are well folded (Figure 4c and Supplementary Figure S2). These spectra overlay each other, thus suggesting that the extra tails at both N- and C-terminal ends do not affect the domain core structure.

Application to the polyketide synthase PpsC from *M. tuberculosis*

Polyketides comprise natural compounds that are essential for the virulence of major human mycobacterial pathogens, namely *M. tuberculosis* (36,37) and other emerging infectious agents. Polyketide biosynthesis is accomplished by polyketide synthases (PKS), which are giant and multifunctional enzymes. PpsC belongs to the family of type I PKS with six domains present on a single polypeptide chain.

Information on the 3D structure of full-length type I PKS is not available and the current data collected worldwide on fatty acid synthases (FAS) and the 6-deoxyerythronolide B synthase (DEBS) from *Saccharopolyspora erythraea*, the most studied modular PKS, only allow rough modeling of their molecular architectures, thus providing only a low-resolution picture. High-resolution structural information is needed to describe at the atomic level each individual domain in order to understand the full-length PKS catalytic machinery. Dedicated software for the analysis of PKS, MAPSI

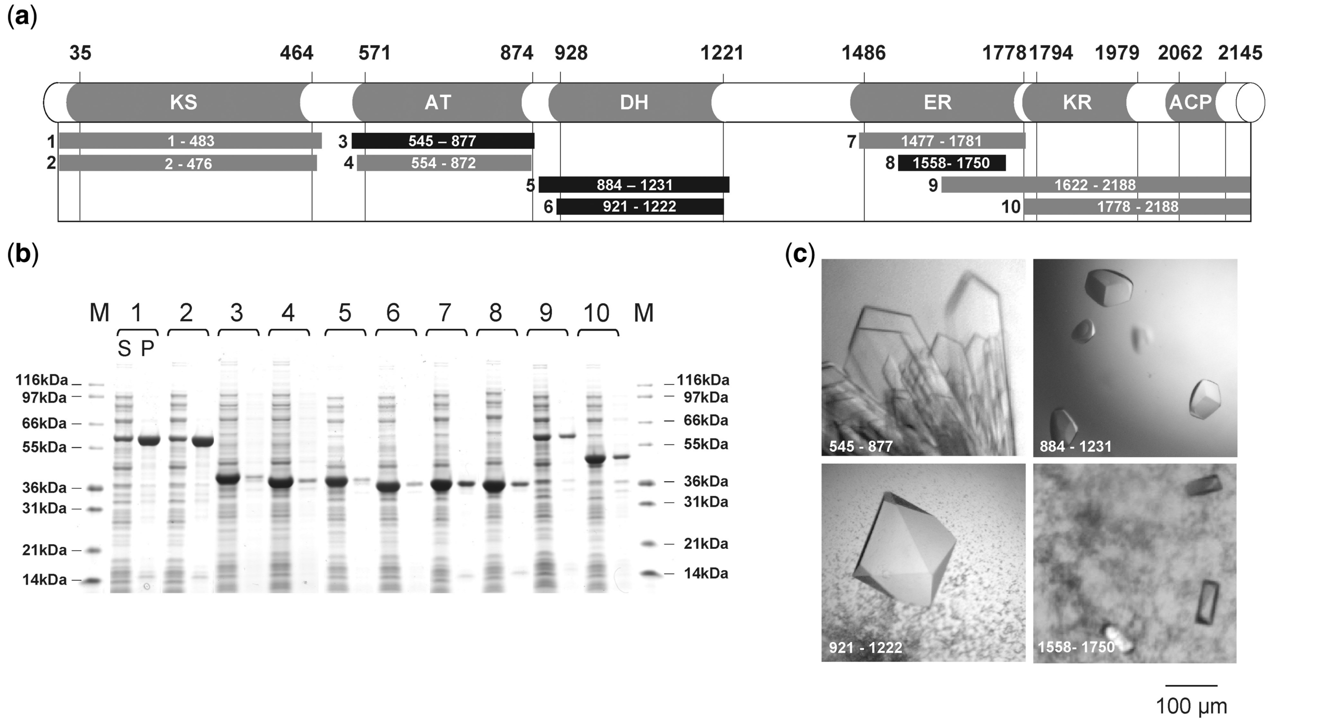

Figure 6. Expression, solubility and crystallization trials of selected PpsC fragments. (**a**) Ten fragments covering the structural domains of PpsC were subcloned from the pTET-GFP 11 plasmid into a N6–HIS pET vector for biochemical and biophysical characterization. (**b**) SDS–PAGE of soluble (S) and pellet (P) fractions of *E. coli* BL21 (DE3) cells expressing the different fragments. (**c**) Pictures of X-ray quality crystals of selected fragments represented as a black rectangle in (a) are shown. The 3D structures of the AT (545–877), DH (921–1222), and truncated ER (1558–1750) domains have been recently determined at 1.8, 2.8, and 2.5 Å resolutions, respectively.

(http://gate.smallsoft.co.kr:8008/pks/mapsitools/index.pl) and SEARCHPKS (http://www.nii.res.in/searchpks.html) were used to identify approximate boundaries for each individual domain. Only in the case of the DH domain, the identification of C-terminal boundaries differs substantially for the methods used in the two approaches (1093 with SEARCHPKS and 1221 with MAPSI).

We used a hydrodynamic point-sink shearing method (38) to create large-size (850–1650 kb) DNA libraries of *Ppsc* fragments. Compared to enzymatic digestion with DNase I, the resulting DNA fragment libraries were tightly distributed in size (data not shown). For each targeted domain, IPCR was used to generate focused sublibraries of fragments, as shown in Figure 5. Initially, phosphorylated forward and reverse IPCR primers were designed to the center of the KS, AT, DH, ER and ACP domains (Supplementary Data 1). Multiple priming sites originating from the high GC content of the *M. tuberculosis* genome lead to PCR amplification failures. To circumvent this problem, 32-mer primers were blasted against the full-length gene to ensure priming at a unique site. An ensemble of 10 fragments covering all six PpsC domains was selected (Figure 6a) and subcloned from a pTET-GFP 11 vector using NdeI and SpeI sites (Supplementary Data 1) into a N6–HIS pET vector. Selected fragments had to satisfy two main criteria: (i) N- and C-terminal boundaries should incorporate the predicted ones and (ii) fragments with different levels of solubility should be considered if possible. SDS–PAGE of the soluble and pellet fractions of *E. coli* BL21 (DE3) cells, expressing the different fragments, were in good agreement with the *in vitro* split-GFP solubility screen (Figure 6b). Crystallization trials have been performed on 6 out of 10 selected constructs centered on the AT, DH and ER domains. Diffraction quality crystals have been obtained for all three domains (Figure 6c). Analysis of the 3D structure of the fragment 921–1222 centered onto the DH domain clearly indicates that the typical double-hotdog fold extends from positions 933 to 1216 (A. Faille *et al.*, manuscript in preparation).

DISCUSSION

Our approach to screening libraries of over a million clones can be of general interest for identifying soluble constructs of ‘recalcitrant’ proteins, as it does not require the use of an automated robotic platform. It simply relies on well-proven technologies capable of effectively eliminating the unwanted constructs, thus reducing the population size of fragments to be analysed. Although *E. coli* cells expressing self-associated DHFR in the presence of TMP somewhat forces the selection of inserted in-frame fragments, we noticed the presence of short antisense peptides devoid of stop codons and in-frame with both ends of DHFR. Following the *in vivo* split-GFP filtration step, the proportion of false positives drops to <3% of the total number of picked clones. The most plausible explanation is that they have been eliminated during gel purification prior to subcloning into the GFP S11 vector.

The IPCR step is crucial as it helps remove the bias originating from the overrepresentation of regions of the protein that never get amplified due to the lack of homology to the IPCR primers. It also permits single amino acid resolution boundary mapping as seen from both p85α and PpsC sublibraries of fragments. For example, in the case of the p85α SH3 domain, two distinct populations of fragments at the junction between amino acid positions 6 and 9 were visible (Figure 3). Detailed examination of the solution structure of the SH3 domain (31) revealed that Tyr6 is the last amino acid residue of the N-terminal tail that connects to a four-residues β-strand central to a triple-stranded antiparallel β-sheet. Cutting into this strand would not only have a destabilizing effect on interactions within the β-sheet, but also with a parallel β-sheet of two strands crossing at right angle. In the case of PpsC, single amino acid truncations also have a dramatic impact on the solubility of the expressed fragments, a result which clearly demonstrates the potential of our approach in an attempt to identify potential candidates for downstream functional and structural applications. To date, only the crystal structure of a fragment covering partially the ER domain has been reported (PDB code 1PQW). Thanks to our approach, we successfully crystallized fragments encompassing the AT, DH and ER domains and solved the X-ray structure of the active AT and DH domains (Alexandre Faille *et al*, manuscripts in preparation).

Our domain trapping strategy is also particularly well adapted to situations, where unstructured regions are essentials for the stability of isolated domains. As seen with p85α and PpsC, it offers a rapid and easy way to identify N- and C-terminal boundaries of soluble fragments often difficult to predict solely by theoretical means. In the near future, we anticipate our approach will facilitate decipher structure–function relationships in mechanistically diverse and complex enzymatic machineries, thus opening the way to atomic level description of active sites and domain–domain interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the staff of synchrotron beamlines ID14-1, ID14-2, ID23-2 and ID29 at the European Synchrotron Radiation Facility (Grenoble, France); Françoise Viala for her help in preparing the figures. We would like to thank Dr Fengli Zhang at National High Magnetic Field Laboratory (Florida, USA) for his help with NMR spectroscopy.

FUNDING

National Institutes of Health's Protein Structure Initiative (grant number 5U54GM074946-4); the French 'Agence Nationale de la Recherche' (ANR-09-BLAN-0298-01). Funding for open access charge: National Institutes of Health's Protein Structure Initiative (grant number 5U54GM074946-4).

Conflict of interest statement. The split-GFP and related intellectual properties are the subject of domestic and foreign patent applications by Los Alamos National Laboratories on behalf of the Department of Energy and LANS, L.L.C.

REFERENCES

1. Pedelacq,J.D., Piltch,E., Liong,E.C., Berendzen,J., Kim,C.Y., Rho,B.S., Park,M.S., Terwilliger,T.C. and Waldo,G.S. (2002) Engineering soluble proteins for structural genomics. *Nat. Biotechnol.*, **20**, 927–932.
2. Pedelacq,J.D., Cabantous,S., Tran,T., Terwilliger,T.C. and Waldo,G.S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.*, **24**, 79–88.
3. Sippl,M.J. (2009) Fold space unlimited. *Curr. Opin. Struct. Biol.*, **19**, 312–320.
4. Prodromou,C., Savva,R. and Driscoll,P.C. (2007) DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug Discov. Today*, **12**, 931–938.
5. Cornvik,T., Dahlroth,S.L., Magnusdottir,A., Herman,M.D., Knaust,R., Ekberg,M. and Nordlund,P. (2005) Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat. Methods*, **2**, 507–509.
6. Reich,S., Puckey,L.H., Cheetham,C.L., Harris,R., Ali,A.A., Bhattacharyya,U., Maclagan,K., Powell,K.A., Prodromou,C., Pearl,L.H. *et al.* (2006) Combinatorial domain hunting: an effective approach for the identification of soluble protein domains adaptable to high-throughput applications. *Protein Sci.*, **15**, 2356–2365.
7. Waldo,G.S., Standish,B.M., Berendzen,J. and Terwilliger,T.C. (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.*, **17**, 691–695.
8. Kawasaki,M. and Inagaki,F. (2001) Random PCR-based screening for soluble domains using green fluorescent protein. *Biochem. Biophys. Res. Commun.*, **280**, 842–844.
9. Yumerefendi,H., Tarendeau,F., Mas,P.J. and Hart,D.J. (2010) ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *J. Struct. Biol.*, **172**, 66–74.
10. Nadal,M., Mas,P.J., Blanco,A.G., Arnan,C., Sola,M., Hart,D.J. and Coll,M. (2010) Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain. *Proc. Natl Acad. Sci. USA*, **107**, 16078–16083.
11. Dyson,M.R., Perera,R.L., Shadbolt,S.P., Biderman,L., Bromek,K., Murzina,N.V. and McCafferty,J. (2008) Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucleic Acids Res.*, **36**, e51.
12. Nakayama,M. and Ohara,O. (2003) A system using convertible vectors for screening soluble recombinant proteins produced in *Escherichia coli* from randomly fragmented cDNAs. *Biochem. Biophys. Res. Commun.*, **312**, 825–830.
13. Zacchi,P., Sblattero,D., Florian,F., Marzari,R. and Bradbury,A.R. (2003) Selecting open reading frames from DNA. *Genome Res.*, **13**, 980–990.
14. Cho,G., Keefe,A.D., Liu,R., Wilson,D.S. and Szostak,J.W. (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J. Mol. Biol.*, **297**, 309–319.
15. Seehaus,T., Breitling,F., Dubel,S., Klewinghaus,I. and Little,M. (1992) A vector for the removal of deletion mutants from antibody libraries. *Gene*, **114**, 235–237.
16. Daugelat,S. and Jacobs,W.R. Jr (1999) The Mycobacterium tuberculosis recA intein can be used in an ORFTRAP to select for open reading frames. *Protein Sci.*, **8**, 644–653.

17. Lutz,S., Fast,W. and Benkovic,S.J. (2002) A universal, vector-based system for nucleic acid reading-frame selection. *Protein Eng.*, **15**, 1025–1030.
18. Gerth,M.L., Patrick,W.M. and Lutz,S. (2004) A second-generation system for unbiased reading frame selection. *Protein Eng. Des. Sel.*, **17**, 595–602.
19. Smith,V.F. and Matthews,C.R. (2001) Testing the role of chain connectivity on the stability and structure of dihydrofolate reductase from *E. coli*: fragment complementation and circular permutation reveal stable, alternatively folded forms. *Protein Sci.*, **10**, 116–128.
20. Cabantous,S., Terwilliger,T.C. and Waldo,G.S. (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.*, **23**, 102–107.
21. Hoskins,R.A., Stapleton,M., George,R.A., Yu,C., Wan,K.H., Carlson,J.W. and Celniker,S.E. (2005) Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res.*, **33**, e185.
22. Cabantous,S. and Waldo,G.S. (2006) *In vivo* and *in vitro* protein solubility assays using split GFP. *Nat. Methods*, **3**, 845–854.
23. Listwan,P., Terwilliger,T.C. and Waldo,G.S. (2009) Automated, high-throughput platform for protein solubility screening using a split-GFP system. *J. Struct. Funct. Genomics*, **10**, 47–55.
24. Marley,J., Lu,M. and Bracken,C. (2001) A method for efficient isotopic labeling of recombinant proteins. *J. Biomol. NMR*, **20**, 71–75.
25. Delaglio,F., Grzesiek,S., Vuister,G.W., Zhu,G., Pfeifer,J. and Bax,A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
26. Wishart,D.S., Bigam,C.G., Yao,J., Abildgaard,F., Dyson,H.J., Oldfield,E., Markley,J.L. and Sykes,B.D. (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, **6**, 135–140.
27. Tarendeau,F., Boudet,J., Guilligay,D., Mas,P.J., Bougault,C.M., Boulo,S., Baudin,F., Ruigrok,R.W., Daigle,N., Ellenberg,J. *et al.* (2007) Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat. Struct. Mol. Biol.*, **14**, 229–233.
28. Tarendeau,F., Crepin,T., Guilligay,D., Ruigrok,R.W., Cusack,S. and Hart,D.J. (2008) Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. *PLoS Pathog.*, **4**, e1000136.
29. Guilligay,D., Tarendeau,F., Resa-Infante,P., Coloma,R., Crepin,T., Sehr,P., Lewis,J., Ruigrok,R.W., Ortin,J., Hart,D.J. *et al.* (2008) The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat. Struct. Mol. Biol.*, **15**, 500–506.
30. Angelini,A., Tosi,T., Mas,P., Acajjaoui,S., Zanotti,G., Terradot,L. and Hart,D.J. (2009) Expression of Helicobacter pylori CagA domains by library-based construct screening. *FEBS J.*, **276**, 816–824.
31. Booker,G.W., Gout,I., Downing,A.K., Driscoll,P.C., Boyd,J., Waterfield,M.D. and Campbell,I.D. (1993) Solution structure and ligand-binding site of the SH3 domain of the p85 alpha subunit of phosphatidylinositol 3-kinase. *Cell*, **73**, 813–822.
32. Musacchio,A., Cantley,L.C. and Harrison,S.C. (1996) Crystal structure of the breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 alpha subunit. *Proc. Natl Acad. Sci. USA*, **93**, 14373–14378.
33. Nolte,R.T., Eck,M.J., Schlessinger,J., Shoelson,S.E. and Harrison,S.C. (1996) Crystal structure of the PI 3-kinase p85 amino-terminal SH2 domain and its phosphopeptide complexes. *Nat. Struct. Biol.*, **3**, 364–374.
34. Siegal,G., Davis,B., Kristensen,S.M., Sankar,A., Linacre,J., Stein,R.C., Panayotou,G., Waterfield,M.D. and Driscoll,P.C. (1998) Solution structure of the C-terminal SH2 domain of the p85 alpha regulatory subunit of phosphoinositide 3-kinase. *J. Mol. Biol.*, **276**, 461–478.
35. Derewenda,Z.S. (2010) Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 604–615.
36. Camacho,L.R., Ensergueix,D., Perez,E., Gicquel,B. and Guilhot,C. (1999) Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. *Mol. Microbiol.*, **34**, 257–267.
37. Cox,J.S., Chen,B., McNeil,M. and Jacobs,W.R. Jr (1999) Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice. *Nature*, **402**, 79–83.
38. Oefner,P.J., Hunicke-Smith,S.P., Chiang,L., Dietrich,F., Mulligan,J. and Davis,R.W. (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.*, **24**, 3879–3886.