Journal of Structural Biology 175 (2011) 159–170

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Expression of protein complexes using multiple *Escherichia coli* protein co-expression systems: A benchmarking study

Didier Busso [a,1], Yoav Peleg [b,1], Tatjana Heidebrecht [c], Christophe Romier [a], Yossi Jacobovitch [b], Ada Dantes [b], Loubna Salim [a], Edouard Troesch [a], Anja Schuetz [d], Udo Heinemann [d,e], Gert E. Folkers [f], Arie Geerlof [g,2], Matthias Wilmanns [g], Andrea Polewacz [h], Claudia Quedenau [h], Konrad Büssow [h,i], Rachel Adamson [j], Elena Blagova [k], Julia Walton [k], Jared L. Cartwright [j], Louise E. Bird [l,3], Raymond J. Owens [l,3], Nick S. Berrow [l,4], Keith S. Wilson [k], Joel L. Sussman [b,m], Anastassis Perrakis [c], Patrick H.N. Celie [c,*]

[a] *Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Institut National de Santé et de Recherche Médicale (Inserm), U964/Centre National deRecherche Scientifique (CNRS), UMR 7104, Université de Strasbourg, 1 rue Laurent Fries, 67404 Illkirch, France*
[b] *The Israel Structural Proteomics Center (ISPC), Faculty of Biochemistry, Weizmann Institute of Science, Rehovot 76100, Israel*
[c] *NKI Protein Facility and Division of Biochemistry, The Netherlands Cancer Institute (NKI), Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands*
[d] *Helmholtz Protein Sample Production Facility, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Str. 10, 13125 Berlin, Germany*
[e] *Institut für Biochemie, Freie Universität Berlin, Takustr. 6, 14095 Berlin, Germany*
[f] *Bijvoet Centre for Biomolecular Research, Utrecht University, NMR Spectroscopy Department, Padualaan 8, 3584 CH Utrecht, The Netherlands*
[g] *EMBL Hamburg Outstation, Notkestrasse 85, 22603 Hamburg, Germany*
[h] *Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany*
[i] *Department of Structural Biology, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany*
[j] *Protein Production Laboratory, Technology Facility, Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK*
[k] *YSBL, Department of Chemistry, Structural Biology Laboratory, University of York, Heslington, York YO10 5DD, UK*
[l] *OPPF, Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK*
[m] *Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel*

ARTICLE INFO

Article history:
Available online 5 March 2011

Keywords:
Escherichia coli
Co-expression
Cloning strategies
Enzyme-free cloning
Gateway™
In-Fusion™
LIC
Restriction-free cloning

ABSTRACT

Escherichia coli (*E. coli*) remains the most commonly used host for recombinant protein expression. It is well known that a variety of experimental factors influence the protein production level as well as the solubility profile of over-expressed proteins. This becomes increasingly important for optimizing production of protein complexes using co-expression strategies. In this study, we focus on the effect of the choice of the expression vector system: by standardizing experimental factors including bacterial strain, cultivation temperature and growth medium composition, we compare the effectiveness of expression technologies used by the partners of the Structural Proteomics in Europe 2 (SPINE2-complexes) consortium. Four different protein complexes, including three binary and one ternary complex, all known to be produced in the soluble form in *E. coli*, are used as the benchmark targets. The respective genes were cloned by each partner into their preferred set of vectors. The resulting constructs were then used for comparative co-expression analysis done in parallel and under identical conditions at a single site. Our data show that multiple strategies can be applied for the expression of protein complexes in high yield. While there is no 'silver bullet' approach that was infallible even for this small test set, our observations are useful as a guideline to delineate co-expression strategies for particular protein complexes.

1. Introduction

Multi-protein complexes are often key-regulators in many cellular processes. These complexes can differ in size, varying from only two or three-components to large multimeric-complexes (Charbonnier et al., 2008; Doucet and Hetzer, 2010; Riccio, 2010). Systems biology data have generated many insights into the different pathways and protein networks at the cellular level (Charbonnier et al., 2008). Within the past decade, results from both *in vivo* and *in vitro* studies have illustrated the importance

* Corresponding author. Fax: +31 20 5121954.
E-mail address: p.celie@nki.nl (P.H.N. Celie).
[1] These authors contributed equally to this work and should be considered co-first authors.
[2] Present address: Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, Germany.
[3] Present address: OPPF-UK, Research Complex at Harwell, R92 Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxford OX11 0FA, UK.
[4] Present address: Protein Expression Unit, Institute for Research in Biomedicine, Parc Cientific de Barcelona, C/Josep Samitier, 1–5, 08028 Barcelona, Spain.

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.03.004

of analyzing the composition and mechanisms of protein assembly to unravel complex biological processes. To obtain the protein assemblies which are the subject of biochemical, biophysical and structural analyses necessary to achieve such mechanistic insight, one can isolate endogenous complexes, either by *in vitro* reconstitution from individually expressed protein components, or by heterologous expression of all components in the same host cell. A large effort has been made in technological developments allowing co-expression of recombinant proteins in both prokaryotes and eukaryotes. Expression in eukaryotic cells, such as *sf9* insect cells or mammalian cell lines, may be favored because of post-translational modifications that are essential for protein function and/or stability and because of the presence of particular chaperone systems that may improve protein folding. An example is the expression of a 400 kDa heterohexameric subcomplex of human TFIID containing two copies of each of the three TAF proteins, which was successfully expressed in insect cells using the baculovirus expression system (Fitzgerald et al., 2006). Despite the advantages of the eukaryotic systems, *Escherichia coli* remains the primary system of choice for expressing protein complexes (Bieniossek et al., 2009; Perrakis and Romier, 2008; Romier et al., 2006; Tan et al., 2005; Tolia and Joshua-Tor, 2006). Expression in *E. coli* has the benefit of obtaining large quantities at low cost and at short time, for either individual proteins or protein complexes. In addition, integration of both DNA-cloning and protein expression technologies in well-established high-throughput platforms allow parallel testing of multiple protein variants, as well as different strains and/or culture conditions (Berrow et al., 2006; Vijayyachandran et al., this issue). Moreover, the absence of particular post-translational modifications (e.g. glycosylation) within the *E. coli* system is sometimes an advantage for X-ray crystallography studies, where non-homogenous protein preparations are likely to have an adverse effect on the success-rate of finding crystallization hits. Co-expression in *E. coli* is a strategy that can often present advantages over *in vitro* reconstitution or re-folding of the individually expressed partners, allowing proper folding of the protein partners and formation of a soluble complex *in vivo*, thus overcoming solubility problems of the individually expressed components (Li et al., 1997; Romier et al., 2006).

Many factors can influence the expression of proteins in *E. coli*, including the bacterial strain used for expression, expression system, growth medium and temperature of induction (Berrow et al., 2006; Graslund et al., 2008). In addition to the factors that may influence expression of individual proteins, the experimental results of protein co-expression are affected by several specific factors. These include the choice of partner, position of the affinity-tag (C- or N-terminal) used for co-purification (Diebold et al., this issue; Fribourg et al., 2001; Romier et al., 2006) and the selection of the protein domains used in the co-expression study (Fribourg et al., 2001).

The selected strategy used for protein co-expression may also have an additional impact. Co-expression can be conducted using either single or multiple constructs. In the case of a single plasmid, this can be either poly-cistronic, (i.e. having a single promoter for multiple genes that are transcribed in the same mRNA) or, alternatively, the plasmid can contain multiple genes, each controlled by a separate promoter (transcribed each in a distinct mRNA). When two or more constructs are co-transformed into a single cell, each vector should at least comprise a different antibiotic selection marker (Perrakis and Romier, 2008; Zeng et al., 2010) and each vector could harbor compatible (i.e. distinct) or incompatible (i.e. similar) replicons (Johnston et al., 2000; Perrakis and Romier, 2008; Velappan et al., 2007; Yang et al., 2001).

In the present study, we conducted a systematic benchmarking study exploring the effect of different co-expression strategies, as reflected by the choice of expression vectors, on the production and solubility of different complexes. Within the SPINE2-complexes consortium, each partner has its own set of preferred, often customized, vectors that are suited for protein co-expression. Therefore, we aimed to perform a systematic analysis of different vectors, which were commonly used at eight SPINE2-complexes consortium partner sites (Division of Biochemistry, The Netherlands Cancer Institute (NKI), Amsterdam; Helmholz Protein Sample Production Facility (PSPF), Berlin/Braunschweig; Structural Biology Unit, EMBL-Hamburg Outstation, Hamburg; Oxford Protein Production Facility (OPPF), Division of Structural Biology, Oxford; The Israel Structural Proteomics Center (ISPC), Weizmann Institute of Science, Rehovot; Integrative Structural Biology Program, Institute of Genetic, Molecular and Cellular biology (IGBMC), Strasbourg; NMR spectroscopy research group, Bijvoet center for Biomolecular Research, Utrecht and the Protein Production Laboratory, Department of Biology, University of York York). To compare the different co-expression systems, four protein complexes (three binary and one ternary) were selected, of which only one protein per complex contained an N-terminal Histidine tag for purification purposes (see Section 2.). Most expression vectors tested were based on the T7 promoter system for transcriptional regulation in combination with *E. coli* strain harboring the DE3 prophage (Studier et al., 1990). DNA cloning into the different expression vectors was performed at each individual partner site and protein co-expression was subsequently performed at one site (NKI, Amsterdam), under standardized experimental parameters and to minimize random variations. Our data show that multiple strategies can be applied for expression of complexes in high yield; there does not appear to be a preferred strategy yielding systematically optimal results for all four tested complexes. This emphasizes the importance of efficient high-throughput expression and purification methods also as the means to explore different strategies for a given problem to efficiently choose the best approach by trial and error.

2. Materials and methods

2.1. Selected complexes

Four protein complexes were selected for benchmarking the different co-expression vectors: (1) human Geminin:Cdt1, a 76.6 kDa trimeric complex with 2:1 stroichiometry (De Marco et al., 2009); (2) human TFIIEα:TFIIEβ, a 82.5 kDa dimeric complex (Jawhari et al., 2006); (3) viral influenza Importin-α5:PB2, a 58.6 kDa dimeric complex (Tarendeau et al., 2007); and (4) human NFYC:NFYB:NFYA, a 32.3 kDa trimeric complex (Romier et al., 2006). Details of the proteins and the selected domains thereof are presented in Table 1.

Original DNA constructs containing the respective genes were gathered and amplified at the NKI and subsequently distributed among SPINE2-complexes partners to be used as a template for re-cloning into the expression vectors of choice. All vectors used by each partner are described below and schematic diagrams with the details for all vectors are presented in Fig. 1.

The co-expression trials were categorized in four groups depending on the expression strategy. Group 1 and 2 comprises those trials for which proteins are expressed from multiple plasmids with either incompatible or compatible origin of replications, respectively. Expression trials from constructs that contain multiple genes under control of a single promoter (poly-cistronic transcript) or under control of separate promoters comprise groups 3 and 4, respectively (Table 2). In some expression trials of the ternary his-NFYC:NFYB:NFYA complex, a combination of strategies is used, e.g. when two plasmids with compatible origin of replications are used and one of these contains two genes for bi-cistronic expression (strategy 2 and 3 combined).

Table 1
Protein complexes used for co-expression experiments.

Complex	Protein components	Residues	Mr (kDa)	N-terminal his-tag	Stoichiometry
Cdt1:Geminin[a]	Geminin	1–209 (fl)	23.6	–	1:2
	Cdt1	158–396	29.4	+	
TFIIEα:TFIIEβ[b]	TFIIEα	1–439 (fl)	49.5	+	1:1
	TFIIEβ	1–291 (fl)	33.0	–	
Importin α5:PB2[c]	Importin α5	66–512	49.7	–	1:1
	PB2	678–759	8.9	+	
NFYA:NFYB:NFYC[d]	NFYA	262–347	10.4	–	1:1:1
	NFYB	51–143	10.8	–	
	NFYC	28–120	11.1	+	

The boundaries of each protein (fragment) used for co-expression studies are specified by the residue numbers. Full length proteins are indicated (fl). Mr: Molecular mass in kDa.
[a] De Marco et al., 2009.
[b] Jawhari et al., 2006.
[c] Tarendeau et al., 2007.
[d] Romier et al., 2006.

2.2. Cloning strategies

Each partner performed the cloning into their selected set of vectors using their favorite cloning technologies, such as: Restriction-based, Gateway™ (Invitrogen, Carlsbad, CA), In-Fusion™ (Clontech, Mountain View, CA), LIC (Ligation Independent Cloning) (Aslanidis and de Jong, 1990; Haun et al., 1992) that may be coupled to EFC (Enzyme-Free Cloning) (de Jong et al., 2006), or RF (Restriction-Free) cloning (van den Ent and Löwe, 2006). The integrity of the open reading frame of each of the target genes was verified by sequencing and constructs were sent to the NKI for comparative co-expression profiling experiments using the different co-expression strategies described above.

2.2.1. Oxford Protein Production Facility (OPPF), Oxford University, UK

The OPPF has a selection of pOPIN-expression vectors that contain different tags and selection markers (Berrow et al., 2007). For the present study, the pOPINF vector has been used. This vector is derived from the pTriEX2 plasmid from Merck–Novagen (Darmstadt, Germany) and has been adapted for any insert to be cloned in frame by In-Fusion™ once the pOPINF vector is linearized by appropriate enzymes (*Kpn*I + *Hind*III).

To create a pOPINF construct for expression of two or three genes from the same vector, primers for PCR were designed so as to add a linker region harboring a RBS between each gene pair. The forward primer for the first gene and the reverse primer for the second gene encompass the sequence required for In-Fusion™ reaction with linearized pOPINF vector. The resulting polycistronic plasmids contain multiple genes expressed under the same promoter but with separate RBS (strategy 3).

2.2.2. Institute of Genetics and Molecular and Cellular Biology (IGBMC) Strasbourg, France

Within the IGBMC a large variety of expression vectors is available, all based on Merck–Novagen plasmids: (1) pnEA-tH and ppEA-tH, based on pET15b; (2) pnEK and ppEK, based on pET28a; (3) pnCS and ppCS, based on pCDF-1b; (4) pHGWA, based on pET22b; (5) pCoGWA, based on pETDuet-1; (6) pCo0GWC, based on pACYCDuet-1 and (7) pCo0GWS based on pCDFDuet-1. For the pn and pp series, single gene cloning was done by restriction-ligation using *Nde*I and *Bam*HI sites. Two single-protein expression constructs can subsequently be concatenated by restriction-ligation with compatible restriction sites resulting in a new expression vector harboring both genes that are either under control of separate promoters (pp series) or both under control of the same promoter (i.e. bi-cistronic; pn series) (Diebold et al., this issue; Romier et al., 2006). The pHGWA vectors allow single cloning by the Gateway™ technology (Busso et al., 2005). The pCo vectors enable cloning of two genes; one by restriction-ligation and one by the Gateway™ cloning technology (Busso et al., in preparation). The constructs that were made for the benchmarking study cover three of the different strategies (strategies 2, 3 and 4).

2.2.3. EMBL Hamburg Unit, Germany

The EMBL Hamburg has developed a large variety of expression vectors. In this study, four vectors have been used: pETM-11, pETM-13, pCDF-11 and pCDF-13. The pETM and the pCDF vectors are based on the pET24d and on the pCDF-1b plasmids from Merck–Novagen, respectively. The pETM-11 and pCDF-11 vectors allow expression of the protein with a N-terminal 6xHis-tag followed by a Tobacco Etch Protease (TEV) cleavage site whereas pETM-13 and pCDF-13 do not add any tag.

In this study, constructs for single protein expression of binary complexes were made by restriction-ligation using *Nco*I and *Xho*I sites. The pET- and pCDF-based vectors contain different resistance markers and origin of replications which renders them suitable for co-expression studies following strategy 2.

2.2.4. ISPC Weizmann Institute, Rehovot, Israel

At the Weizmann Institute, two vectors were used: (1) pET28-TEVH, a modified pET28a vector (Merck–Novagen) (Peleg and Unger, 2008), which allows single protein expression with a N-terminal 6×His-tag followed by a TEV cleavage site; and (2) the pACYCDuet-1 vector (Merck–Novagen) which can be used for expression of two proteins, each under control of an independent T7 promoter. The cloning was done either by In-Fusion™ (Clontech, Mountain View, CA), using PCR-linerarized vectors (Benoit et al., 2006), or by Restriction Free (RF) cloning (Unger et al., 2010; van den Ent and Löwe, 2006).

Cloning into the pET28-TEVH was done by insertion of the target genes immediately following the TEV cleavage site and at the *Hind*III site. For constructs expressing two genes, the cloning was done into the pACYCDuet-1 vector. Cloning into the first expression cassette was performed immediately following the *Bam*HI site and at the *Hind*III site. The gene integration into the second expression cassette was performed at the *Nde*I-*Xho*I sites. The resulting constructs made were suitable for conducting the benchmarking study following strategies 2 and 4.

2.2.5. Department of NMR Spectroscopy, Utrecht University, The Netherlands

The Utrecht University has constructed the pLIC and pLICHIS vectors (de Jong et al., 2006) based on the pET15b plasmid (Merck-Novagen) and the pCDFLICHIS vectors adapted from the pCDF-1b plasmid (Merck–Novagen). Those vectors have been elaborated in order to perform cloning for single protein expression by

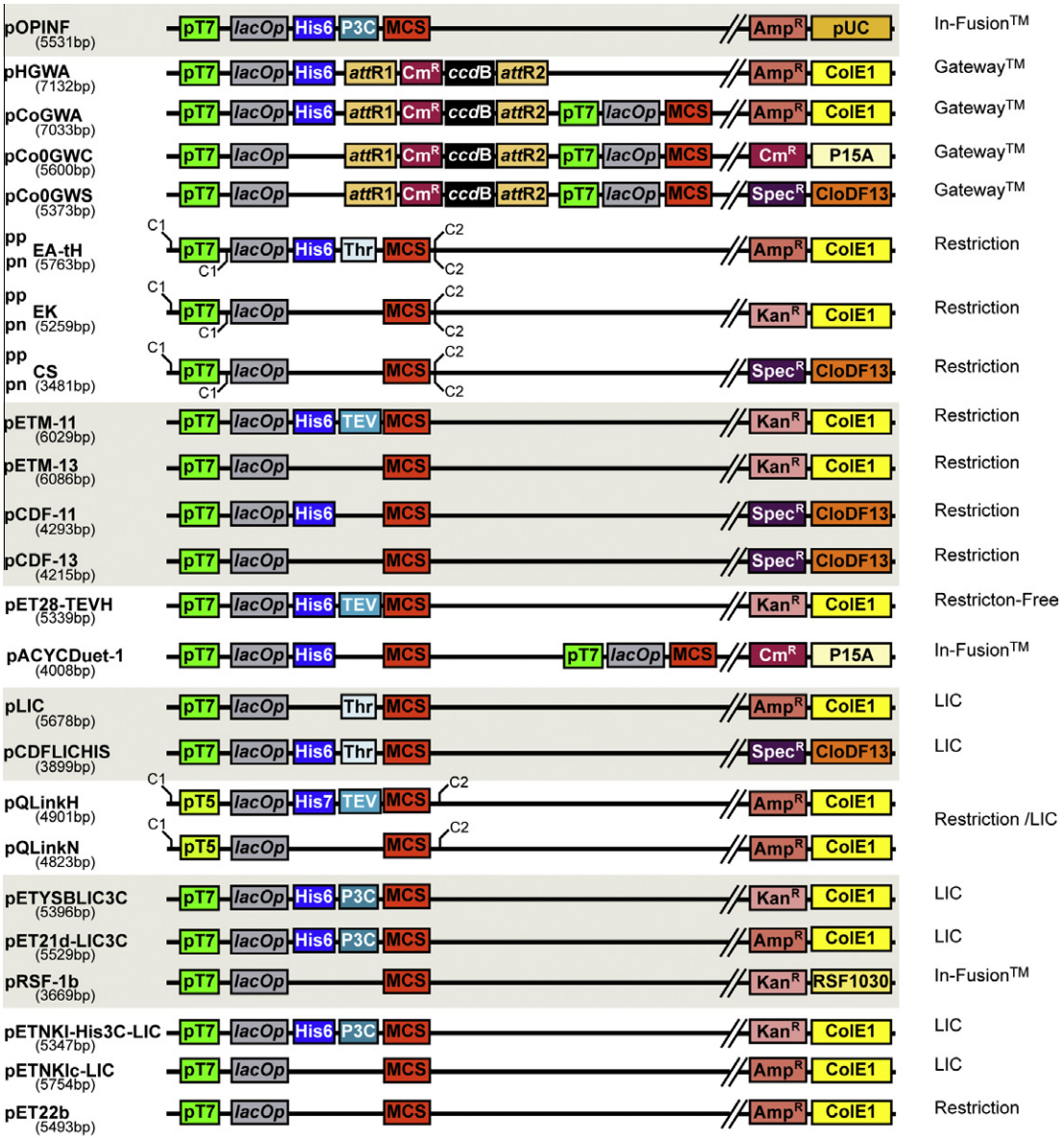

Fig. 1. Schematic representation of the different vectors that were used in the benchmarking study. The name, vector size (in base-pairs) and cloning strategy (right column) are given for all vectors used in this study. Promoters are displayed in green colors (pT7: T7 promoter; pT5: T5 promoter). Sequence encoding tags or recognition cleavage sites are in blue (His6: 6x-histidine tag; His7: 7x-histidine tag; P3C: recognition sequence for HRV-3C protease; TEV: recognition sequence for TEV protease; Thr: recognition sequence for thrombin protease). Resistance gene markers are shown in pink-ish colors (Amp^R: resistance marker for ampicillin; Cm^R: resistance marker for chloramphenicol; Kan^R: resistance marker for kanamycine; $Spec^R$: resistance marker for spectinomycin). Origin of replication are in yellow colors (pUC: 50–100 copies; ColE1: 40 copies; P15A: 10–12 copies; CloDF13: 20–40 copies; RSF1030:>100 copies). *Att*R: recombination site for Gateway cloning during LR reaction. C1 and C2: location of the restriction sites used for concatenation (see Section 2.2.2 and 2.2.6 for details). *ccdB*: gene encoding for the selective marker CcdB protein. *lacOp*: lactose operator sequence. MCS: multiple cloning site.

Ligation Independent Cloning (LIC) coupled with Enzyme Free Cloning (EFC) (de Jong et al., 2006). Briefly, the gene of interest was amplified by performing two PCR amplifications: first with a LIC extended forward primer and a reverse primer; and second with a forward primer and a LIC extended reverse primer. Both PCR products were then mixed, melted and re-annealed, before adding them to *Sac*II-digested and T4-treated vectors. The constructs can be used for co-expression studies according to strategy 2.

2.2.6. Protein Sample Production Facility (PSPF) MDC, Berlin and HZI Braunschweig, Germany

The PSPF has designed a set of vectors named pQLink (Scheich et al., 2007) that are derived from the pQE-2 plasmid (Qiagen, Hilden, Germany). For this study, the gene encoding the protein that contains a 7xHis-tag was cloned by restriction-ligation into pQLinkH vector digested with *Bam*HI and *Not*I. Genes expressing untagged protein were cloned into pQLinkN in the same way. An exception was made for the NFY subunits, which were cloned into the pQLinkH (NFYB) and pQLinkN (NFYB and NFYC) using LIC technology. The pQLink vectors present specific restriction sites (*Pac*I and *Swa*I), which can be used to concatenate single protein expression constructs by LIC (Scheich et al., 2007). The final constructs used in this study were poly-cistronic plasmids containing two or three genes under control of independent promoters (strategy 4).

2.2.7. Protein Production Laboratory (PPL) at the University of York, UK

The PPL used three vectors for this study: (1) pET-YSBLIC3C, a vector based on the pET28a plasmid from Merck–Novagen (Fogg

Table 2
Co-expression strategies used for complex expression.

Partner	Co-expression number	Co-expression vectors	Co-expression strategy[a]
OPPF (Oxford)	1	pOPINF	3
IGBMC (Strasbourg)	2	pCoGWA	4
	3	pHGWA + pCo0GWS	2
	4	pHGWA + pCo0GWS	2/4
	5	pCo0GWC + pCoGWA	2/4
	6	pCo0GWS + pCoGWA	2/4
	7	pnEA-tH + pnCS	2
	8	ppEA-tH + ppCS	2
	9	pnEA-tH + pnEK + pnCS	2
	10	ppEA-tH + ppEK + ppCS	2
	11	pnCS + pnEA-tH	2/3
	12	ppCS + ppEA-tH	2/4
	13	pnEA-tH	3
	14	ppEA-tH	4
EMBL (Hamburg)	15	pETM-11 + pCDF-13	2
	16	pETM-13 + pCDF-11	2
ISPC (Rehovot)	17	pET28-TEVH + pACYCDuet-1	2
	18	pACYCDuet-1	4
	19	pET28-TEVH + pACYCDuet-1	2/4
Bijvoet center (Utrecht)	20	pLICHIS + pCDFLICHIS	2
PSPF (Berlin/ Braunschweig)	21	pQLINK	4
Protein production Laboratory (PPL) (York)	22	pET-YSBLIC3C	4
	23	pET21d_LIC3C	4
	24	pRSF + pET21d_LIC3C	2/4
NKI (Amsterdam)	25	pETNKI-his3C-LIC + pET22b	1
	26	pETNKI-his3C-LIC + pETNKIc-LIC	1

[a] Strategy 1 represents co-expression using multiple plasmids with incompatible origin of replications. Strategy 2 represents co-expression using multiple plasmids with compatible origin of replications. Strategy 3 represents co-expression using a single vector containing multiple genes with one promoter/vector (poly-cistronic). Strategy 4 represents co-expression using a single vector containing multiple genes with one promoter/gene.

and Wilkinson, 2008), (2) the pET21d_LIC3C, based on the pET21d vector (Merck–Novagen), and (3) the pRSF-1b vector (Merck–Novagen). The pET-YSBLIC3C and the pET21d_LIC3C have been adapted for LIC cloning and harbor a sequence encoding an N-terminal 6xHis tag followed by a HRV-3C cleavage site.

For expression of the binary complexes, pET-YSBLIC3C and pET21d-LIC3C vector were digested with *Bse*RI and subsequently treated with T4 DNA polymerase. The two genes were amplified by PCR using appropriate LIC primers and were treated with T4 DNA polymerase and subsequently mixed with the LIC-Duet Minimal Adaptor (Merck–Novagen) in order to have an additional copy of the T7 promoter for the second gene. The resulting plasmids contain two genes, each under control of their own T7 promoter (strategy 4).

For the NFY ternary complex, two out of the three genes were cloned into pET21d_LIC3C vector as described above and the third gene was cloned into the *Nco*I digested pRSF-1b vector using In-Fusion™ technology (Clontech, Mountain View, CA). The resulting plasmid comprises a compatible origin of replication when used in combination with pET21d_LIC3C.

2.2.8. Protein Facility and Division of Biochemistry, NKI Amsterdam, The Netherlands

At the NKI, two LIC-based vectors have been constructed: The pETNKI-His3C-LIC vector based on pET28a (Merck–Novagen), and the pETNKIc-LIC vector based on pET46-Ek/LIC (Merck–Novagen). For single protein expression of the 6×His-tagged protein, PCR product of each corresponding gene was treated with T4 DNA polymerase and annealed with the pETNKI-His3C-LIC vector digested by *Kpn*I and treated with T4 DNA polymerase (Luna-Vargas et al., submitted for publication). The gene required for binary complexes expression was cloned into pETNKIc-LIC, digested with *AaRI* and treated with T4 DNA polymerase, or inserted into pET22b vector (Merck–Novagen) by restriction-ligation methods. The resulting plasmids present incompatible origin of replication (strategy 1).

2.3. General protein expression procedures

Protocols for protein expression and purification for each complex were provided by the partner who supplied the initial constructs. Protocols were standardized between the expression of the different complexes with respect to expression volume (50 ml cultures) and *E. coli* strain used. Plasmids containing the selected genes were transformed into BL21(DE3) cells and plated on LB-agar plates with appropriate antibiotics. When using multiple vectors in a single transformation, colonies containing the plasmids were selected on plate with antibiotics for each vector. Single colonies were grown over-night in 2 ml LB medium (+antibiotics) at 37 °C and were added to 48 ml of fresh LB medium (+antibiotics) the next day. Transformed cells were grown at 37 °C until optical cell density at 595 nm (OD595) reached 0.6–0.8. The temperature was then lowered and protein expression induced by addition of IPTG for a defined duration according to partner protocols (see supplementary information for details).

After cultivation, cells were harvested by centrifugation (3000*g* for 15 min) and the weight of the cell pellet was measured. For each gram of cell mass, 5 ml of the appropriate lysis buffer was added (see supplementary information). Cells were disrupted by sonication and cell debris and the insoluble fraction was removed by centrifugation at 10,000*g* for 30 min at 4 °C. The soluble fraction for each extract was loaded onto 250 µl of immobilized metal affinity resin charged with Ni^{2+} (Merck–Novagen) and incubated for 60 min (except for Geminin:Cdt1 complex that was incubated for 15 min) at 4 °C. Beads were washed 3 times with 4 column volumes (CV) (except for Geminin:Cdt1 complex that was washed once with 10 CV) of defined buffer (see supplementary information for details) and proteins were eluted with 2 CV of elution buffer (see supplementary information).

2.4. Analysis and quantification of protein expression levels

The volume of elution fractions was determined and protein concentrations were calculated by measuring OD at 280 nm on a nanodrop spectrophotometer (Thermofisher Scientific Inc.) to get an estimate of protein quantities. Samples of soluble- and elution fractions were loaded on 15% SDS–PAGE gels, except for analysis of the his-NFYC:NFYB:NFYA complex for which a 20% SDS–PAGE gel was used to get optimal separation between these three proteins. Gels were stained with Coomassie brilliant blue for protein visualization. In parallel, the same samples were loaded on a LabChip GXII automated capillary gel electrophoresis system (Caliper LifeSciences, Hopkinton, MA, USA) according to the manufacturer protocol. Protein bands are visualized using a fluorescent Gel-Dye solution provided with the LabChip GXII kit and the instrument optics detect the laser-induced fluorescent protein signal. System software automatically analyzes the data and determines protein size and concentration relative to a ladder and a marker calibration standard.

Sample 26 (his-PB2:Importin α5, Table 5), and samples 11 and 12 (both his-NFYC:NFYB:NFYA, Table 6) were not used in the

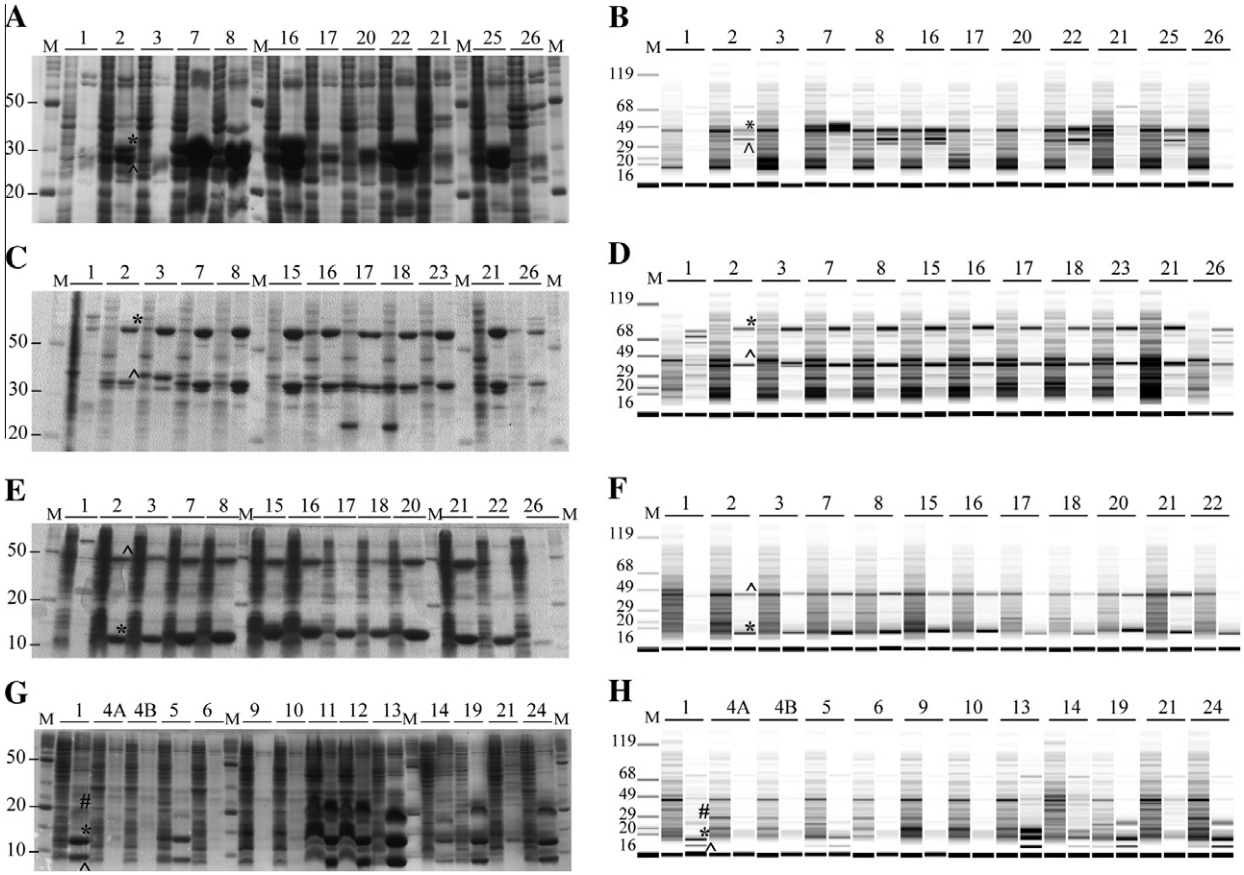

Fig. 2. SDS–PAGE analysis of co-expression trials. Samples were loaded on standard SDS–PAGE gels and protein bands were visualized by coomassie brilliant-blue staining (A, C, E, G). For each co-expression trial 5 μl of the soluble fraction (always the left lane) and 15 μl of the elution fraction (right lane) were loaded on gel. The same samples were also loaded onto the Labchip GXII system (Caliper LifeSciences) according to the manufacturer protocol (B, D, F, H), with the soluble fraction always in the left lane and the elution fraction in the right lane. (A) and (B), his-Cdt1 (*) and Geminin (^); (C) and (D), his-TFIIEα (*) and TFIIEβ (^); (E) and (F), his-PB2 (*) and Importin α5 (^); (G) and (H), his-NFYC (*), NFYB (^) and NFYA (#). Numbers on top of the gels represent the co-expression experiment number as described in Table 2. Molecular weight markers (M) are indicated.

LabChip GXII system. In order to quantify these proteins, Image J image integration software (Abramoff et al., 2004) was used to calculate protein concentrations from SDS–PAGE gels. The intensity of the protein bands was determined and compared to that of other samples of the his-PB2:Importin α5 and his-NFYC:NFYB:NFYA complex, for which protein concentrations were quantified using the LabChip GXII software.

3. Results

3.1. Selection of protein complexes

The selection of four protein complexes was made based on some simple criteria. First, a stable protein complex had to be previously demonstrated either by co-expression or by expression of the individual partners separately followed by *in vitro* reconstitution of the complex. Second, it should be possible to purify the whole complex in a single step when only one of the proteins is fused to a his-tag. Third, we aimed for both binary and ternary complexes as well as complexes with different stoichiometry. Fourth, the components were chosen to cover a broad range of molecular weights. Finally, we only considered complexes available in SPINE-2-complexes participants' labs to avoid intellectual property issues or other practical complications. Based on these requirements, Cdt1:Geminin complex (De Marco et al., 2009), TFIIEα:TFIIEβ complex (Jawhari et al., 2006), Importin α5:PB2 complex (Tarendeau et al., 2007) and NFYA:NFYB:NFYC complex (Romier et al., 2006) were selected (Table 1).

3.2. Co-expression of his-Cdt1:Geminin complex (1:2)

Co-expression and purification of both full length and truncated his-Cdt1:Geminin complexes have been described previously (Lee et al., 2004; De Marco et al., 2009). Structural analysis revealed that the his-Cdt1:Geminin complex used in our study, exists predominantly as a [Cdt1:2xGeminin] heterotrimer, which can dimerise in solution to give rise to a heterohexamer, a mechanism we showed to be important for cellular function (De Marco et al., 2009). In this study, the complex could be expressed and purified in eight out of twelve expression trials (Fig. 2A and B and Table 3). His-Cdt1:Geminin production varies considerably between different vector systems. In the majority of experiments his-Cdt1 and Geminin are co-expressed, which is consistent with the notion that expression of soluble his-Cdt1 is strongly dependent on the presence of its binding partner Geminin. This is exemplified when comparing experiment 25 (pETNKI-his3C-LIC + pET22b) and experiment 26 (pETNKI-his3C-LIC + pETNKI-his3C-LIC), which only differ in the vector used for Geminin expression. In trial 25, a decent amount of complex is produced, whereas virtually no complex is obtained from experiment 26.

The best results were obtained in co-expression trial 22 (pET-YSBLIC3C), 16 (pCDF-11 + pETM-13) and 8 (ppEA-tH + ppCS). Trial 22 is part of strategy 4, where both genes are under the control

Table 3
His-Cdt1:Geminin co-expression results.

Co-expression number	Partner	Construct	Co-expression strategy	µg his-Cdt1/ g cells[a]	µg Geminin/ g cells[a]	Molar Ratio his-Cdt1:Geminin
1	OPPF	pOPINF his-Cdt1:Geminin	3	11	8.0	1 : 0.9
2	IGBMC	pCoGWA his-Cdt1:Geminin	4	112	83	1 : 0.9
3	IGBMC	pHGWA-his-Cdt1 + pCo0GWS-Geminin	2	<1	<1	n.d.[b]
7	IGBMC	pnEA-tH-his-Cdt1 + pnCS-Geminin	2	982	327	1 : 0.4
8	IGBMC	ppEA-tH-his-Cdt1 + ppCS-Geminin	2	488	439	1 : 1.1
16	EMBL	pCDF-11-his-Cdt1 + pETM-13-Geminin	2	581	576	1 : 1.2
17	ISPC	pET28-his-Cdt1 + pACYCDuet-Geminin	2	56	38	1 : 0.8
20	Bijvoet center	pCDFLIC-his-Cdt1 + pLIC-Geminin	2	50	6	1 : 0.2
22	PPL	pET-YSBLIC3C-hisCdt1:Geminin	4	572	675	1 : 1.5
21	PSPF	pQLink-Geminin:his-Cdt1	4	30	16	1 : 0.7
25	NKI	pETNKI-his3C-LIC-his-Cdt1 + pET22b-Geminin	1	280	223	1 : 1.0
26	NKI	pETNKI-his3C-LIC-his-Cdt1 + pETNKlc-LIC-Geminin	1	14	9	1 : 0.6

[a] Protein concentrations were determined from the LabChip GXII gelelectrophoresis experiments (Fig. 2B).
[b] n.d.: Not determined.

of individual promoters on the same vector, whereas in the other two trials both proteins are expressed from individual vectors with compatible origins (strategy 2). For the majority of purified complexes, the relative amount of Geminin compared to that of purified his-Cdt1 is (slightly) lower than expected from the 1:2 stoichiometry (Table 3 and Fig. 3A), suggesting a minor excess of his-Cdt1, which is not in complex with Geminin.

Together, these data suggest that the level of Geminin expression is a key determinant for obtaining large quantities of the complex, and the choice of vectors could make the difference between large amounts of soluble protein or little or no protein at all.

3.3. Co-expression of his-TFIIEα:TFIIEβ complex (1:1)

The general human RNA polymerase II transcription factor TFIIE is composed of two subunits, TFIIEα and TFIIEβ and the complex can be purified to homogeneity as a heterodimeric complex in a 1:1 stoichiometry (Jawhari et al., 2006). For all of the twelve constructs tested in our study, soluble his-TFIIEα:TFIIEβ complex could be purified in a 1:1 (±0.3) ratio (Figs. 2C and 3B and Table 4). There is an approximate 10-fold difference in expression amount between the most- and least- optimal vector systems (trial 8 and 1, respectively). Six experiments (3, 7, 8, 15, 16 and 17) that all include co-expression from multiple plasmids, produce large amounts of soluble complex. In addition, co-expression from a single vector (experiment 18, 21 and 23) also resulted in production of relatively large amount of complex (Table 2, strategies 3 and 4). In conclusion, most – but not all – of the vectors studied are suited for production of the his-TFIIEα:TFIIEβ complex, and expression yields would differ considerably depending on the exact choice.

3.4. Co-expression of his-PB2:Importin α5 complex (1:1)

Co-expression of the PB2:Importin-α5 complex has not been reported before, however the heterodimeric complex could be formed after *in vitro* reconstitution of individually-expressed proteins, and the crystal structure of this complex was determined (Tarendeau et al., 2007). In our experiments, both proteins – and in particular his-PB2 – can be produced in the soluble form (Figs. 2E, F and 3C). Although the complex can be purified in most of the trials, an excess of his-PB2 is always co-purified (typically more than 5-fold compared to Importin α5), indicating that the expression levels of Importin α5 is the key-determinant in obtaining purified complex in high yields. This suggests that formation of the complex itself is not essential for enhancing expression of the individual components. This is compatible with previous data showing that single expression of each protein results in multi-milligram quantities of soluble protein (Tarendeau et al., 2007).

When using (untagged) Importin α5 as measure for the amount of complex that is formed, vector systems 7, 8, 15 (all part of strategy 2) and 21 (strategy 4) prove to be most successful. The pQLink system (trial 21) appears to be the optimal system for expression of this complex, since it produces the highest amount of Importin α5, although the difference is only 2-fold compared to systems 7, 8 and 15. At least two systems (1, 26) practically fail to produce this complex, while most other systems result in considerably lower yields when compared with the best trials.

3.5. Co-expression of his-NFYC:NFYB:NFYA complex (1:1:1)

The ternary complex of the transcription factor NFYC:NFYB:NFYA can be co-expressed and purified (Romier et al., 2006). NFYC forms a tight dimer with the NFYB subunit (Romier et al., 2003), a prerequisite for NFYA association. The resulting trimer binds with high specificity and affinity to 5′-CCAAT-3′ DNA motifs in the promoter region. Expression of ternary complexes is in general more challenging compared to binary complexes because many co-expression vectors only allow expression of two genes and co-expression from multiple vectors requires incompatible origin of replications and three different antibiotics. In our experiments we were able to express the ternary complex from eight different construct combinations (Table 6, Fig. 2G and H). It should be noted that for all samples, the amount of NFYB that is quantified appears to be lower than expected based on the intensity of the bands on SDS–PAGE (Fig. 2G and H). NFYB migrates beyond the 14 kDa lower-limit range of the LabChip GXII gel and calculation of the protein content may therefore be underestimated. The highest amount of complex was obtained from the poly-cistronic pnEA-tH construct, where only one promoter is present (trial 13). Remarkably, when additional promoter sites are inserted before each individual gene (ppEA-tH; trial 14), expression is significantly reduced. However, this difference is not observed when NFYA is expressed from a separate vector (either pn-CS or pp-CS) and the NFY subunits B and C are both expressed from a single construct under control of one (trial 11) or two (trial 12) promoter sites (see also Diebold et al., submitted for publication). In this case, only about half of the vectors tried produced any ternary complex at all, while within this half the yields would differ within at least one order of magnitude.

4. Discussion

We examined the effect of different co-expression systems on the production of soluble protein complexes in *E. coli*. The

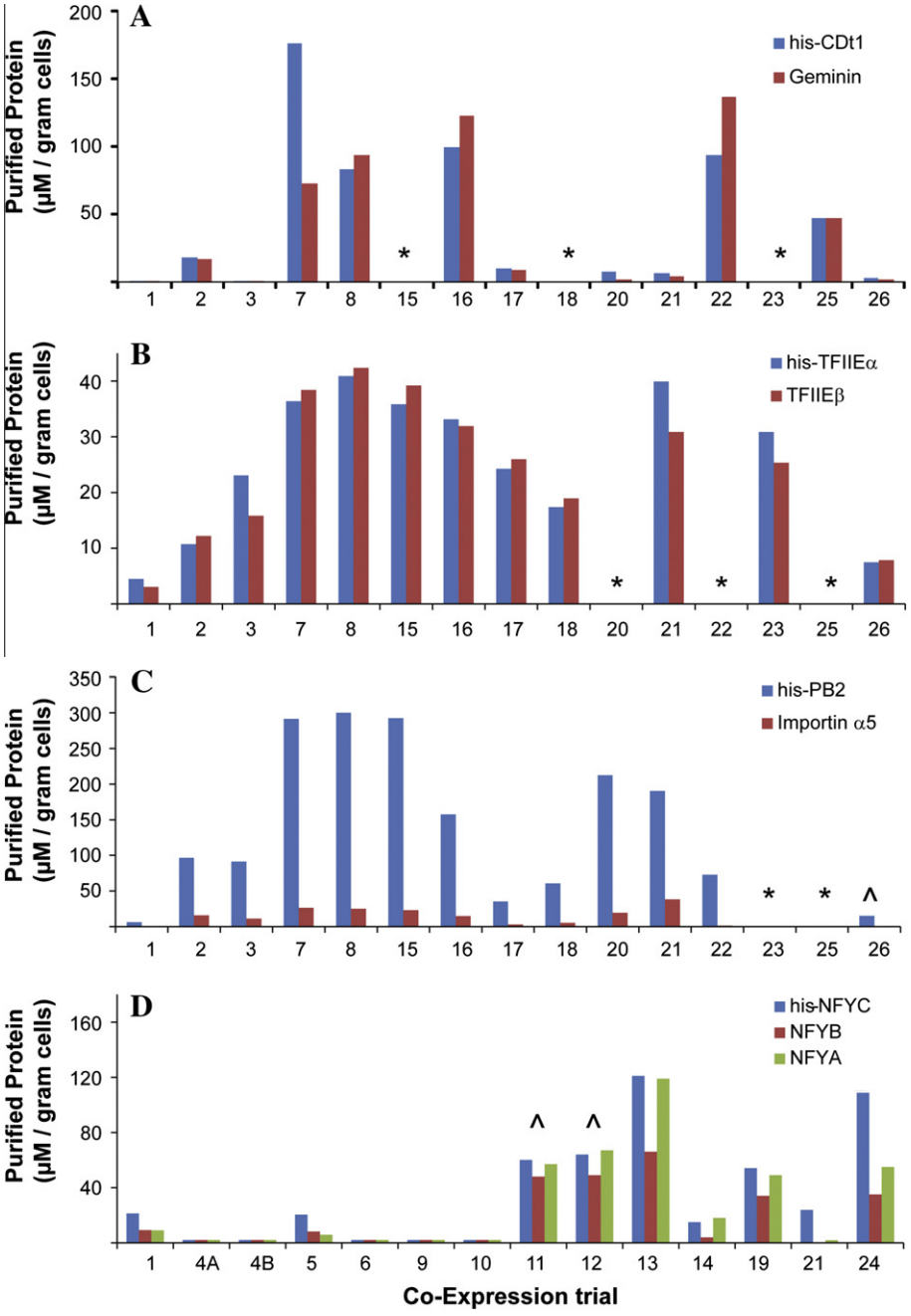

Fig. 3. Comparison of co-expression profiles. The amount of purified protein (in µM per gram cell mass) that is calculated from the LabChip GXII data is plotted for each experiment. (A) his-Cdt1:Geminin; (B) his-TFIIEα:TFIIEβ; (C) his-PB2:Importin α5; (D) his-NFYC:NFYB:NFYA. His-tagged proteins are always represented by the blue bars. Numbers below the different columns represent the co-expression experiment number as described Table 2. *This vector (combination) was not tried for this particular complex. ^Protein concentrations were calculated using Image J software as described in Section 2.4.

collection of vectors used provides a nice coverage of expression strategies that are currently used in major European labs. Almost all sites use customized commercial vectors, modified and optimized for a particular strategy. In addition, many alternative methods for restriction-based cloning are implemented, including LIC, Enzyme-Free, Restriction-Free, In-Fusion™ and Gateway™ cloning. These technologies are beneficial for high-throughput cloning as they bypass the screening for suitable restriction sites. Indeed, many structural genomics centers have successfully applied these methods for high-throughput protein production (Eschenfeldt et al., 2010; Joachimiak, 2009; Savitsky et al., 2010; Xiao et al., 2010).

Comparing the results obtained for the four complexes, different co-expression profiles can be observed (Fig. 3). Production of His-Cdt1:Geminin shows a rather 'black-and-white' pattern; either the complex is virtually not produced at all, or the complex is expressed in decent amounts with relatively small – but still appreciable – differences in expression levels between vector systems (Fig. 3A). This can be explained by our observation that expression of soluble his-Cdt1 relies on the expression of Geminin. The expression profile of the ternary his-NFYC:NFYB:NFYA complex also shows significant variation between the different co-expression systems, indicative for the challenges that may arise during production of multi-protein complexes containing more than two components. First, protein expression from three individual constructs (e.g. experiments 9 and 10) may be complicated if not all plasmids are amplified to the same level which may hamper equal production levels, especially if the correct folding of one

Table 4
His-TFIIEα:TFIIEβ co-expression results.

Co-expression number	Partner	Construct	Co-expression strategy	μg his-TFIIEα/ g cells[a]	μg TFIIEβ/g cells[a]	Molar Ratio his-TFIIEα:TFIIEβ
1	Oxford	pOPINF his-TFIIEα:TFIIEβ	3	67	30	1 : 0.7
2	IGBMC	pCoGWA his-TFIIEα:TFIIEβ	4	108	81	1 : 1.1
3	IGBMC	pHGWA-his-TFIIEα + pCo0GWS-TFIIEβ	2	232	104	1 : 0.7
7	IGBMC	pnEA-tH-his-TFIIEα + pnCS-TFIIEβ	2	366	254	1 : 1.1
8	IGBMC	ppEA-tH-his-TFIIEα + ppCS-TFIIEβ	2	411	280	1 : 1.0
15	EMBL	pETM-11-his-TFIIEα + pCDF-13-TFIIEβ	2	361	259	1 : 1.1
16	EMBL	pCDF-11-his-TFIIEα + pETM-13-TFIIEβ	2	334	211	1 : 1.0
17	ISPC	pET28-his-TFIIEα + pACYCDuet-TFIIEβ	2	244	172	1 : 1.1
18	ISPC	pACYCDuet-his-TFIIEα:TFIIEβ	4	175	125	1 : 1.1
23	PPL	pET21d_LIC3C-his-TFIIEα:TFIIEβ	4	311	167	1 : 0.8
21	PSPF	pQLink-his-TFIIEα:TFIIEβ	4	402	204	1 : 0.8
26	NKI	pETNKI-his3C-LIC-his-TFIIEα + pETNKIc LIC-TFIIEβ	1	105	73	1 : 1.1

[a] Protein concentrations were determined from the LabChip GXII gelelectrophoresis experiments (Fig. 2D).

Table 5
his-PB2:Importin α5 co-expression results.

Co-expression number	Partner	Construct	Co-expression Strategy	μg his-PB2/g cells[a]	μg Importin-α5/ g cells[a]	Molar ratio his-PB2:Importin α5
1	OPPF	pOPINF his-PB2:Importin α5	3	12	5	1 : 0.1
2	IGBMC	pCoGWA his-PB2:Importin α5	4	187	149	1 : 0.2
3	IGBMC	pHGWA-his-PB2 + pCo0GWS-Importin α5	2	168	99	1 : 0.1
7	IGBMC	pnEA-tH-his-PB2 + pnCS-Importin α5	2	624	273	1 : 0.1
8	IGBMC	ppEA-tH-his-PB2 + ppCS-Importin α5	2	521	212	1 : 0.1
15	EMBL	pETM-11-his-PB2 + pCDF-13-Importin α5	2	626	240	1 : 0.1
16	EMBL	pCDF-11-his-PB2 + pETM-13-Importin α5	2	337	153	1 : 0.1
17	ISPC	pET28-his-PB2 + pACYCDuet-Importin α5	2	179	17	1 : < 0.1
18	ISPC	pACYCDuet-his-PB2:Importin α5	4	121	51	1 : 0.1
20	Bijvoet Center	pCDFLIC-his-PB2 + pLIC-Importin α5	2	412	183	1 : 0.1
21	PSPF	pQLink-his-PB2:Importin α5	4	408	398	1 : 0.2
22	PPL	pET-YSBLIC3C-his-PB2:Importin α5	4	141	11	1 : < 0.1
26	NKI	pETNKI-his3C-LIC-his-PB2 + pETNKIc-LIC-Importin α5	1	25[b]	<0.1[b]	n.d.[c]

[a] Protein concentrations were determined from the LabChip GXII gelelectrophoresis experiments (Fig. 2F) unless stated otherwise.
[b] Intensity of protein band on SDS–PAGE was determined using Image J software and the protein concentration was calculated using corresponding bands that have been quantified using the Labchip GXII software (see Section 2.4).
[c] n.d.: Not determined.

subunit is a prerequisite for correct folding of the complex. Second, for poli-cystronic expression from a single construct, the order in which the genes are aligned could be an important factor (compare trials 5 and 6) (Diebold et al., submitted for publication) and third, there is a sequential order in complex assembly: the binary NFYB:NFYC complex should be formed before NFYA can associate to form a ternary assembly (Romier et al., 2006).

For both his-TFIIEα:TFIIEβ and his-PB2:Importin α5 a more robust expression profile is observed. In these cases, expression of the individual components is less dependent on the presence of their respective partners as each of the components can be individually expressed as soluble proteins in *E. coli* (Jawhari et al., 2006; Tarendeau et al., 2007). For both his-TFIIEα:TFIIEβ and his-PB2:Importin α5, the amount of isolated complex varies only up to 3-fold between the 'top-8' most producing constructs, illustrating that multiple strategies can be used for expression of these soluble complexes.

From the results that we obtained for the different complexes, there is no particular strategy that clearly stands out from the others. A valid comparison between the multi-vector systems with either incompatible (strategy 1) or compatible (strategy 2) cannot be made due to the limited number of experiments performed according to strategy 1 (only trials 25 and 26). The majority of trials are based on strategy 2, which therefore dominate the profiles and produce a bias towards this strategy, which nevertheless, appears quite successful for all the binary complexes (Fig. 3A–C). Comparison between co-expression from multiple-gene per vector constructs, comprising either one promoter per vector (strategy 3) or one promoter per gene (strategy 4) is also limited because strategy 3 is significantly under-represented, in particular for expression of the binary complexes. Vectors 21, 22 and 23 (strategy 4) performed quite well, as they all appear at least once in the top-5 of the co-expression profiles of a particular complex.

For the expression of the ternary NFY complex, our data are even less conclusive, as a combination of different strategies (i.e. 2 and 3 or 2 and 4) is used in a substantial number of trials. However, one remarkable observation is that neither of the expressions using three separate plasmids (experiments 9 and 10) produced soluble complex. This could either be because co-expression from two separate vectors is more effective due to plasmid stability/amplification issues or that multi-gene constructs expressing both NFYC and NFYB (trial 1, 5, 11, 12, 13, 14 and 21) enable more efficient formation of the NFYB:NFYC pre-complex, maybe at the translation level. Additionally, co-existence of three expression vectors in the same cell, harboring each a different antibiotic resistance, may cause a substantial increase in burden on the cell, compared to trials were only two expression vectors are being used.

Although it is tempting to search for the best co-expression system within our collection, this should be done with caution. At least three strategies appear most promising: Number 7 and 11

Table 6
his-NFYC:NFYB:NFYA co-expression results.

Co-expression experiment	Partner	Construct	Co-expression strategy	μg his-NFYC/ g cells[a]	μg NFYB/ g cells[a]	μg NFYA/g cells[a]	Molar ratio his-NFYC:NFYB:NFYA
1	OPPF	pOPINF his-NFYC:NFYA:NFYB	3	50	22	19	1:0.4:0.4
4A	IGBMC	pHGWA-his-NFYC + pCo0GWS-NFYB:NFYA	2/4	<1	<1	<1	n.d.[c]
4B	IGBMC	pHGWA-his-NFYC + pCo0GWS-NFYA:NFYB	2/4	<1	<1	<1	n.d.
5	IGBMC	pCoGWA-his-NFYC:NFYB + pCo0GWC-NFYA	2/4	41	17	12	1:0.4:0.3
6	IGBMC	pCoGWA-his-NFYC:NFYA + pCo0GWS-NFYB	2/4	<1	<1	<1	n.d.
9	IGBMC	pnEK-NFYA + pnCS-NFYB + pnEA-tH-his-NFYC	2	<1	<1	<1	n.d.
10	IGBMC	ppEK-NFYA + ppCS-NFYB + ppEA-tH-his-NFYC	2	<1	<1	<1	n.d.
11	IGBMC	pnCS-NFYA + pnEA-tH-NFYB:his-NFYC	2/3	121[b]	109[b]	107[b]	1:0.8:0.9
12	IGBMC	ppCS-NFYA + ppEA-tH-NFYB:his-NFYC	2/4	129[b]	111[b]	125[b]	1:0.8:1.0
13	IGBMC	pnEA-tH-his-NFYC:NFYA:NFYB	3	284	150	260	1:0.3:1.2
14	IGBMC	ppEA-tH-his-NFYC:NFYA:NFYB	4	31	18	35	1:0.6:0.9
19	ISPC	pET28 his-NFYC + pACYCDuet-NFYB:NFYA	2/4	121	74	103	1 : 0.6 : 0.9
21	PSPF	pQLink-his-NFYC:NFYA:NFYB	4	56	<1	4.2	1 :<0.1 : 0.1
24	PPL	pET21d_LIC3C-his-NFYC:NFYA + pRSF-NFYB	2/4	128	79	120	1 : 0.6 : 0.5

[a] Protein concentrations were determined from the LabChip GXII gelelectrophoresis experiments (Fig. 2H) unless stated otherwise.
[b] Intensity of protein band on SDS–PAGE was determined using Image J software and the protein concentration was calculated using corresponding bands that have been quantified using the Labchip GXII software (see Section 2.4).
[c] n.d.: Not determined.

(both pnEA-tH + pnCS), 8 and 12 (both ppEA-tH + ppCS) and 16 (pCDF-11 + pETM-13) all perform above average for at least three complexes. Interestingly, all three systems comprise a combination of a pCDF-Ib-based and a pET-based vector. Within these combinations, each vector contains a different origin of replication with similar copy numbers (CloDF13 and ColE1, respectively) and a different resistance marker. However, some of the other systems prove at least successful for two complexes (nr. 17 and 19; pET28b + pACYCDuet-1 and nr. 21; pQLink). Also, systems that were less frequently used in our trials show good results (e.g. nr. 15, pCDF-13 + pETM-11; nr. 22, pET-YSBLIC3C and 23, pET21d_LIC3C).

Although many factors contribute to the level of protein expression, we aimed to only vary the expression vectors and keep all other variables constant in our experiments (cell-type, temperature, growth medium etc.), to get the best possible comparison of the different strategies and how these are affected by the exact choice of vectors. It should be noted that for some of the constructs that have been expressed at NKI, a higher yield of protein complex was obtained at the respective partner site (data not shown). However, these results were sometimes obtained with modified conditions. Therefore the results obtained in this study provide a better basis for comparison of the data but may not necessarily reflect the optimal expression conditions for each complex, while emphasizing the need for optimization. In addition, some constructs that were designed by the partners but were not included in our series, proved more successful when tested at the respective site. For instance, when two different pOPIN vectors with incompatible origin of replication were used, much higher levels of co-expression were seen compared to that of our results with the pOPINF vector (data not shown).

Experimental variations like *E. coli* strain, growth temperature and culture media should also be applied to decipher the optimal conditions for obtaining a particular protein complex. In addition, there are of course many other factors that could contribute to improved complex formation that are beyond the scope of our study, including bioinformatics analysis for selection of stable protein fragments (He et al., 2009 and references therein; Pirovano and Heringa, 2010 and references therein), the choice of solubility- and/or affinity tags and usage of codon-optimized genes (Burgess-Brown et al., 2008; Welch et al., 2009). An important factor in protein expression that may be underestimated is the ribosomal binding site (RBS). Recently, a thermodynamic model for the prediction of translation initiation as a marker for protein expression efficiency has been proposed (Salis et al., 2009). The model is based on the RBS sequence and takes into account the energy associated with hybridization of the start codon to the initiating tRNA anticodon loop (3′-UAC-5′), the energy penalty for non-optimal spacing between the 16S rRNA binding site and the start codon and energies related to mRNA (un)folding of secondary structure elements. For most vectors containing an N-terminal tag, the RBS sequence is already optimized and does not alter upon insertion of the gene. However, for expression of non-tagged proteins the initial codons of the mRNA transcript could influence translation initiation and hence protein expression and complex formation. This can be exemplified by the pETNKIc-LIC vector, which is used in trial 26. This vector shows virtually no expression of Geminin (Fig. 2A and B) and Importin α5 (Fig. 2E and F) and according to the thermodynamic calculation (http://voigtlab.ucsf.edu/software/) (Salis et al., 2009), the translation initiation from this construct is rather poor due to sub-optimal RBS sequence. Therefore, improved LIC vectors have been designed at NKI to overcome this problem (Luna-Vargas et al., submitted for publication).

Together our results indicate that different co-expression vectors and strategies can be successfully applied for production of protein complexes in high yield. Selection of a particular system depends on the preferred cloning strategy and available information about the complex. For instance, when the individual components can be expressed as single, soluble proteins, the selection of vector system may be less important compared to that of complexes of which the partners are mutually dependent on their expression. In the latter situation, creation of different constructs

will increase the chance of finding a suitable system that may give rise to high-yield complex formation in *E. coli*. The lack of a clear conclusion, even when using a rather limited test set, emphasizes the need to implement efficient high-throughput trial and error cloning and expression testing strategies. However, the current diversity of choices in cloning methods between laboratories discourages the use of many vectors and strategies, since each group of vectors would require different PCR products. A likely challenge for the centers in the INSTRUCT initiative for Structural Biology in Europe (or the PSI centers in the US) will be to streamline cloning strategies to a highly divergent set of expression vectors, to offer truly high throughput trials across many vector systems and co-expression strategies.

Acknowledgments

We would like to thank Darren Hart and Frank Tarendeau for providing his-PB2 and Importin α5 template DNA and Arnaud Poterszman for providing plasmids encoding TFIIE subunits. We would also like to thank Matthieu Stierlé for performing preliminary expression tests and Bernat Blasco for assistance during the co-expression trials. We thank Janett Tischer, Silke Kurths, and Ingrid Berger for excellent technical assistance. We are grateful to Yves Sergeant from Caliper LifeSciences for giving us the opportunity to use the LabChip GXII and Cathleen Salomo for assistance with running the samples on the LabChip GXII.

This research was supported by the European Commission Sixth Framework Research and Technological Development Program 'SPINE2-COMPLEXES' Project, under contract No. 031220. The Protein Sample Production Facility (PSPF) is funded by the Helmholtz Association of German Research Centres. The NKI Protein Facility is funded by NKI and a NWO-Groot grant from The Netherlands Scientific Organisation (NWO). The Israel Structural Proteomics Center (ISPC) is supported by the Israel Ministry of Science, Culture, and Sport, the Divadol Foundation and the Neuman Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2011.03.004.

References

Abramoff, M.D., Magelhaes, P.J., Ram, S.J., 2004. Image processing with image. J. Biophotonics Int. 11, 36–42.

Aslanidis, C., de Jong, P.J., 1990. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. 18, 6069–6074.

Benoit, R.M., Wilhelm, R.N., Scherer-Becker, D., Ostermeier, C., 2006. An improved method for fast, robust, and seamless integration of DNA fragments into multiple plasmids. Protein Expr. Purif. 45, 66–71.

Berrow, N.S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., et al., 2007. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res. 35, e45.

Berrow, N.S., Bussow, K., Coutard, B., Diprose, J., Ekberg, M., et al., 2006. Recombinant protein expression and solubility screening in *Escherichia coli*: a comparative study. Acta Crystallogr. D Biol. Crystallogr. 62, 1218–1226.

Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., et al., 2009. Automated unrestricted multigene recombineering for multiprotein complex production. Nat. Methods 6, 447–450.

Burgess-Brown, N.A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., et al., 2008. Codon optimization can improve expression of human genes in *Escherichia coli*: a multi-gene study. Protein Expr. Purif. 59, 94–102.

Busso, D., Delagoutte-Busso, B., Moras, D., 2005. Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in *Escherichia coli*. Anal. Biochem. 343, 313–321.

Charbonnier, S., Gallego, O., Gavin, A.C., 2008. The social network of a cell: recent advances in interactome mapping. Biotechnol Ann. Rev. 14, 1–28.

de Jong, R.N., Daniëls, M.A., Kaptein, R., Folkers, G.E., 2006. Enzyme free cloning for high throughput gene cloning and expression. J. Struct. Func. Genom. 7, 109–118.

De Marco, V., Gillespie, P.J., Li, A., Karantzelis, N., Christodoulou, E., et al., 2009. Quaternary structure of the human Cdt1-Geminin complex regulates DNA replication licensing. Proc. Natl Acad. Sci. USA 106, 19807–19812.

Diebold, M.L., Fribourg, S., Koch, M., Metzger, T., Romier, C., Deciphering correct strategies for multiprotein complex assembly by co-expression: application to complexes as large as the histone octamer, J. Struct. Biol. Submitted for publication.

Doucet, C.M., Hetzer, M.W., 2010. Nuclear pore biogenesis into an intact nuclear envelope. Chromosoma 119, 469–477.

Eschenfeldt, W.H., Maltseva, N., Stols, L., Donnelly, M.I., Gu, M., et al., 2010. Cleavable C-terminal His-tag vectors for structure determination. J. Struct. Func. Genom. 11 (1), 31–39.

Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., et al., 2006. Protein complex expression by using multigene baculoviral vectors. Nat. Methods 3, 1021–1032.

Fogg, M.J., Wilkinson, A.J., 2008. Higher-throughput approaches to crystallization and crystal structure determination. Biochem. Soc. Trans. 36, 771–775.

Fribourg, S., Romier, C., Werten, S., Gangloff, Y.G., Poterszman, A., et al., 2001. Dissecting the interaction network of multiprotein complexes by pairwise coexpression of subunits in *E. coli*. J. Mol. Biol. 306, 363–373.

Graslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., et al., 2008. Protein production and purification. Nat. Methods 5, 135–146.

Haun, R.S., Serventi, I.M., Moss, J., 1992. Rapid, reliable ligation-independent cloning of PCR products using modified plasmid vectors. Biotechniques 13, 515–518.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., et al., 2009. Predicting intrinsic disorder in proteins: an overview. Cell Res. 19, 929–949.

Jawhari, A., Uhring, M., De Carlo, S., Crucifix, C., Tocchini-Valentini, G., et al., 2006. Structure and oligomeric state of human transcription factor TFIIE. EMBO Rep. 7, 500–505.

Joachimiak, A., 2009. High-throughput crystallography for structural genomics. Curr. Opin. Struct. Biol. 19, 573–584.

Johnston, K., Clements, A., Venkataramani, R.N., Trievel, R.C., Marmorstein, R., 2000. Coexpression of proteins in bacteria using T7-based expression plasmids: expression of heteromeric cell-cycle and transcriptional regulatory complexes. Protein Expr. Purif. 20, 435–443.

Li, C., Schwabe, J.W., Banayo, E., Evans, R.M., 1997. Coexpression of nuclear receptor partners increases their solubility and biological activities. Proc. Natl. Acad. Sci. USA 94, 2278–2283.

Lee, C., Hong, B., Choi, J.M., Kim, Y., Watanabe, S., et al., 2004. Structural basis for inhibition of the replication licensing factor Cdt1 by geminin. Nature 430, 913–917.

Luna-Vargas, M.P.A., Christodoulou, E., Alfieri, A., Dijk, W.J.V., Stadnik, M., et al., J. Struct. Biol. Submitted for publication.

Peleg, Y., Unger, T., 2008. Application of high-throughput methodologies to the expression of recombinant proteins in *E. coli*. Methods Mol. Biol. 426, 197–208.

Perrakis, A., Romier, C., 2008. Assembly of protein complexes by coexpression in prokaryotic and eukaryotic hosts: an overview. Methods Mol. Biol. 426, 247–256.

Pirovano, W., Heringa, J., 2010. Protein secondary structure prediction. Methods Mol. Biol. 609, 327–348.

Riccio, A., 2010. Dynamic epigenetic regulation in neurons: enzymes, stimuli and signaling pathways. Nat. Neurosci. 13, 1330–1337.

Romier C., Cocchiarella, F., Mantovani, R., Moras, D., 2003. The NF-YB/NF-YC Structure Gives Insight into DNA Binding and Transcription Regulation by CCAAT Factor NF-Y. 278, 1336–45.

Romier, C., Ben Jelloul, M., Albeck, S., Buchwald, G., Busso, D., et al., 2006. Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. Acta Crystallogr. D Biol. Crystallogr. 62, 1232–1242.

Salis, H.M., Mirsky, E.A., Voigt, C.A., 2009. Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. 27, 946–950.

Savitsky, P., Bray, J., Cooper, C.D., Marsden, B.D., Mahajan, P., et al., 2010. High-throughput production of human proteins for crystallization: the SGC experience. J. Struct. Biol. 172, 3–13.

Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U., Büssow, K., 2007. Vectors for co-expression of an unrestricted number of proteins. Nucleic Acids Res. 35, e43.

Studier, F.W., Rosenberg, A.H., Dunn, J.J., Dubendorff, J.W., 1990. Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol. 185, 60–89.

Tan, S., Kern, R.C., Selleck, W., 2005. The pST44 polycistronic expression system for producing protein complexes in *Escherichia coli*. Protein Expr. Purif. 40, 385–395.

Tarendeau, F., Boudet, J., Guilligay, D., Mas, P.J., Bougault, C.M., 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.

Tolia, N.H., Joshua-Tor, L., 2006. Strategies for protein coexpression in *Escherichia coli*. Nat. Methods 3, 55–64.

Unger, T., Jacobovitch, Y., Dantes, A., Bernheim, R., Peleg, Y., 2010. Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. J. Struct. Biol. 172, 34–44.

van den Ent, F., Löwe, J., 2006. RF cloning: a restriction-free method for inserting target genes into plasmids. J. Biochem. Biophys. Methods 67, 67–74.

Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P., Bradbury, A.R., 2007. Plasmid incompatibility: more compatible than previously thought? Protein Eng. Des. Sel. 20, 309–313.

Vijayyachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., et al. Robots, pipelines, polyproteins: Enabling multiprotein expression in prokaryotic and eukaryotic cells, J. Struct. Biol. Submitted for publication.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., et al., 2009. Design parameters to control synthetic gene expression in *Escherichia coli*. PLoS ONE 14 (4), e7002.

Xiao, R., Anderson, S., Aramini, J., Belote, R., Buchwald, W.A., et al., 2010. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. J. Struct. Biol. 172, 21–33.

Yang, W., Zhang, L., Lu, Z., Tao, W., Zhai, Z., 2001. A new method for protein coexpression in Escherichia coli using two incompatible plasmids. Protein Expr. Purif. 22, 472–478.

Zeng, J., Zhang, L., Li, Y., Wang, Y., Wang, M., et al., 2010. Over-producing soluble protein complex and validating protein-protein interaction through a new bacterial co-expression system. Protein Expr. Purif. 69, 47–53.