A Scalable Screening of *E. coli* Strains for Recombinant Protein Expression

Luana G. Morão[a¶], Lívia R. Manzine[a¶], Angélica Luana C. Barra[a,b], Lívia Oliveira D. Clementino[a], Raíssa Gutierrez[a], Carsten Wrenger[c], Igor Polikarpov[a] and Alessandro S. Nascimento[a]*.

a. Pólo TerRa, São Carlos Institute of Physics, University of São Paulo. Av. João Dagnone, 1100, Jd. Santa Angelina, 13563-120. São Carlos, SP. Brazil.

b. Institute of Biochemistry and Molecular Biology, Laboratory for Structural Biology of Infection and Inflammation, University of Hamburg, c/o DESY, Build. 22a. Notkestr. 85, 22603 Hamburg, Germany.

c. Unit for Drug Discovery, Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, Av. Prof. Lineu Prestes 1374, 05508-000 São Paulo-SP, Brazil.

¶ These authors contributed equally to this work.

* Corresponding author: Phone: +55-16-3373-8709. Email: asnascimento@ifsc.usp.br.

Abstract

Structural biology projects are highly dependent on the large-scale expression of soluble protein and, for this purpose, heterologous expression using bacteria or yeast as host systems are usually employed. In this scenario, some of the parameters to be optimized include (*i*) those related to the protein construct, such as the use of a fusion protein, the choice for an N-terminus fusion/tag or a C-terminus fusion/tag; (*ii*) those related to the expression stage, such as the concentration and selection of inducer agent and temperature expression and (*iii*) the choice of the host system, which includes the selection of a prokaryotic or eukaryotic cell and the adoption of a strain. The optimization of some of the parameters related to protein expression, stage (*ii*), is straightforward. On the other hand, the determination of the most suitable parameters related to protein construction requires a new cycle of gene cloning, while the optimization of the host cell is less straightforward. Here, we evaluated a scalable approach for the screening of host cells for protein expression in a structural biology pipeline. We evaluated six *Escherichia coli* strains looking for the best yield in a soluble protein expressed using the same strategy for protein construction and gene cloning. For the proteins used in this experiment, the Arctic Express (DE3) strain resulted in a better yield of soluble protein. We propose that a screening of host cell/strain is feasible, even for smaller laboratories and the experiment as proposed can easily be scalable to a high-throughput approach.

Keywords: Recombinant protein expression, *E. coli* strains, scalable protein expression.

1. Introduction

Many drug discovery projects rely on the structure determination of the drug targets. Indeed, to the best of our knowledge, the so-called 'structure-based rational design' of drug candidates dates from 1976, a year before the Protein Data Bank (PDB) was founded, when Beddell and coworkers designed a series of compounds to fit the human haemoglobin site [1]. Typical structural biology pipelines require pure and soluble protein in high amounts, which is usually achieved by heterologous expression in *E. coli* [2,3] or in yeast cells, such as *Pichia pastoris* or *Saccharomyces cerevisae*, or even in fungus, such as *Aspergillus niger* (e.g., [4]).

E. coli is the host cell of choice in structural biology efforts, due to the ease in handling, rapid doubling time, and very well-established protocols for expression. In our own group experience, protein expression in *E. coli* for high-throughput (HT) pipelines resulted in yields up to 946 mg of protein expressed per liter of culture, with an average of about 60-70 $mg.L^{-1}$ for enzymes of the glycoside hydrolase family [5]. Although this pipeline works very well for most cases, some 'difficult cases' require additional efforts, that include new cloning strategies, such as the addition of fusion tags or even change of tags from C-terminus to N-terminus and vice versa; the exchange of expression vector, suiting of the expression protocol or replacing the host cell.

In this context, several HT strategies have been developed for the parallel cloning of genes of interest in different constructions, which explore different fusion proteins, different tags, and the location of the tag [6–8]. Also, protocols for increasing the yield of soluble expression of target proteins are now abundant in the literature (e.g., [3]). Finally, a myriad of commercial *E. coli* strains is available for efficient heterologous protein expression, promising high yield and enhanced solubility, especially for insoluble protein, making the choice for a particular strain a non-straightforward task.

Here we show the results for a scalable parallel screen for the expression of difficult cases using four different strains. The assay can be readily scalable for an HT format and also include more strains. We show that the screening for expression efficiency in different strains can be a feasible and rapid way to rescue target genes that failed to be soluble expressed in the first HT approach, avoiding going back to the cloning stage of the structural biology pipeline.

2. Materials and methods

All the genes tested in this work were cloned in the pETTRXA-1a/LIC using the strategy previously described by Camilo & Polikarpov [5]. Besides, for a particular target (target **T16**) the pET-NESG vector was used. The pET-Trx1A-LIC is a ligation independent cloning-based vector where the target gene is expressed with an N-terminal His-tag fused to thioredoxin and a cleavage site for the Tobacco Etch Virus (TEV) protease.

2.1 Gene Cloning

The LIC cloning protocol using pETTRXA-1a/LIC vector was performed as described previously [5]. Briefly, the plasmid was linearized by PCR using the following oligonucleotides: Fw- 5´ TGGCGCCCTGAAAATAAAG; Rv- 5´CCGCGTCGGGTCAC and, for targets, genomic DNAs were used as templates in PCR amplification using Phusion High-fidelity DNA Polymerase (New England Biolabs). PCR products were individually treated for 16 h at 37 °C with 20 units of DpnI enzyme (New England Biolabs). Following gel extraction using Wizard SV Gel and PCR CleanUp System (Promega), 500 ng of purified and linearized vector and 200 ng of each specific gene were treated with T4 DNA polymerase enzyme (Fermentas) in the presence of 2.5 mM dTTP (vector) and 2.5 mM dATP (gene) for 30 min at 22 °C. Annealing of LIC vectors and insert was performed after incubation of 3 µL of T4 Polymerase-treated PCR fragments with 1 µL of T4 Polymerase-treated vector at 25 °C for 30 min. Then, the mixture was used for *E. coli* DH5α competent cells transformation and positive clone selection achieved in culture plates containing LB agar medium with 50 µg mL^{-1} kanamycin after overnight incubation at 37 °C.

2.2 Protein Expression

Sixteen target genes were cloned following the procedures above. The resulting vectors were used for *E. coli* expression in five DE3 strains: Lemo21, BL21, pT-GroE, Rosetta-gami 2, and Arctic Express. Transformations were performed in 24-well-LB-agar plates containing: kanamycin (50 µg mL^{-1}) and chloramphenicol (34 ug mL^{-1}) for Lemo21, pTGroE, and Rosetta-gami 2 cells; kanamycin (50 µg mL^{-1}) and gentamycin (20 µg mL^{-1}) for Arctic Express. Colonies were inoculated in DW96 preculture plates containing 1 mL of LB medium supplied with cells specific antibiotics, sealed, and incubated overnight at 37 °C with shaking. Then, 100 µL of preculture was inoculated in 5 DW24 plates containing 4 mL of LB with specific antibiotics (plus 0.5 mM rhamnose for Lemo21 cells) and grown at 37 °C under 150 rpm agitation during 5h. Afterward, aliquots were collected and IPTG induction was performed using 0.4 mM final concentration for Lemo21 and 1 mM for the other strains. The temperature was reduced to 18 °C (pTGroE and Rosetta-Gami 2), 30 °C (Lemo21), or 11 °C (Arctic Express), and expression was conducted for 16 h at 150 rpm. In parallel, 100 µL of pre-culture was inoculated in 5 DW24 plates containing 4 mL of ZYP5052 [9] auto-induction medium, and the expression was performed as LB-conditions without IPTG. The plates were centrifuged for 10 min at 2000xg and each pellet was resuspended with 1 mL of lysis buffer (Tris–HCl 50 mM, NaCl 300 mM, lysozyme 0.25 mg mL^{-1}, 0.1 mM PMSF, at pH 8). Cells were lysed by freezing at -80 °C overnight and thawing at 17 °C for 1 h under agitation. DNAseI (Sigma) was added at 10 µg mL^{-1} final concentration following incubation at 17° C for 15 min. The lysate purification was carried out in an automated platform Freedom EVO 200 (Tecan) using Ni Sepharose 6 Fast Flow Resin (GE) in 96-well Receiver Plates 20 µm (Macherey–Nagel). The lysate passed through the resin by vacuum filtration followed by resin washing with 1 mL of buffer A (Tris–HCl 50 mM, NaCl 300 mM, pH 8) and buffer B (Tris-HCl 50 mM, NaCl 300 mM, imidazole 20 mM, pH 8) for removal of unspecific and unbounded proteins. Elution of interest proteins was performed with 150 µL of buffer C (Tris-HCl 50 mM, NaCl 300 mM, imidazole 300 mM, pH 8) and samples were analyzed by gel electrophoresis for expression protein screening evaluation. The expression yield of the soluble proteins expressed was assessed by Coomassie-stained SDS-PAGE 15%.

3. Results and Discussion

Sixteen 'difficult cases' genes were identified in our typical protein expression pipeline. In this pipeline, the genes are cloned in the pETTRXA-1a/LIC vector and expressed in *E. coli* Rosetta 2 (DE3) strain. In 13 out of the 16 cases, no soluble protein expression could be observed, or the expression yield was very low for a structural biology pipeline. These targets include bacterial (*Staphylococcus aureus and Enterococcus faecalis, Mycobacterium tuberculosis*) enzymes, as well as plasmodial enzymes (*Plasmodium falciparum* and *Plasmodium vivax*). The genes were initially synthesized with codon optimization for *E. coli* expression. The only exception was the *E. faecalis* genes, in which the genes were cloned directly from the genomic DNA. All the constructs present a molecular weight among 30-50 kDa and some of them are under-represented in the PDB, suggesting that they are difficult cases for protein expression in the structural biology pipeline in general. A summary of the genes selected for the screening of soluble expression in *E. coli* strains is shown in Table 1.

The protein purification procedures, as described in the Materials and Methods section, are assisted by liquid-handling robots, making it very easy to scale the assay for an HT format by increasing the number of targets or the number of strains tested. The procedure followed the same HT strategy previously established [5]. The actual bottleneck was the protein expression since every *E. coli* strain has its typical resistance marker and its optimal temperature for expression. In our assay, the growing temperature was set to 37 °C for all strains and different temperatures were used in the induction stage, by allocating each strain in a different shaker. Although this step requires increased human effort, it is easily scalable to include more strains, in particular, if the additional strains can be used in the same expression temperatures, or even for more genes, since they can be used in the same plate in each shaker. After the protein expression step, the samples were collected in a single plate for the purification procedure. The process can be further optimized for an HT format, but it becomes clear that the current format is readily scalable for additional genes/strains.

After the purification, the expression levels obtained were compared by SDS-PAGE analysis. Figure 1 shows the electrophoresis analysis obtained for the expression of the target genes and the results are also summarized in Table 2. Targets **T1**, **T2** and **T7** are orthologous enzymes from *E. faecalis*, *S. aureus*, and *P. falciparum*, respectively, and

all of them showed negligible soluble expression in *E. coli* Rosetta 2 strain. When tested for the expression in Rosetta Gami 2, pT-GroE, LEMO21, and Arctic Express, we observed soluble recovery of **T1** in Rosetta Gami 2, pT-GroE, Lemo21, and Arctic Express. Additionally, **T2** was recovered in minimal amounts in Rosetta Gami 2 and Arctic Express, while **T7** was recovered in higher amounts in Arctic Express and small amounts in the other strains tested.

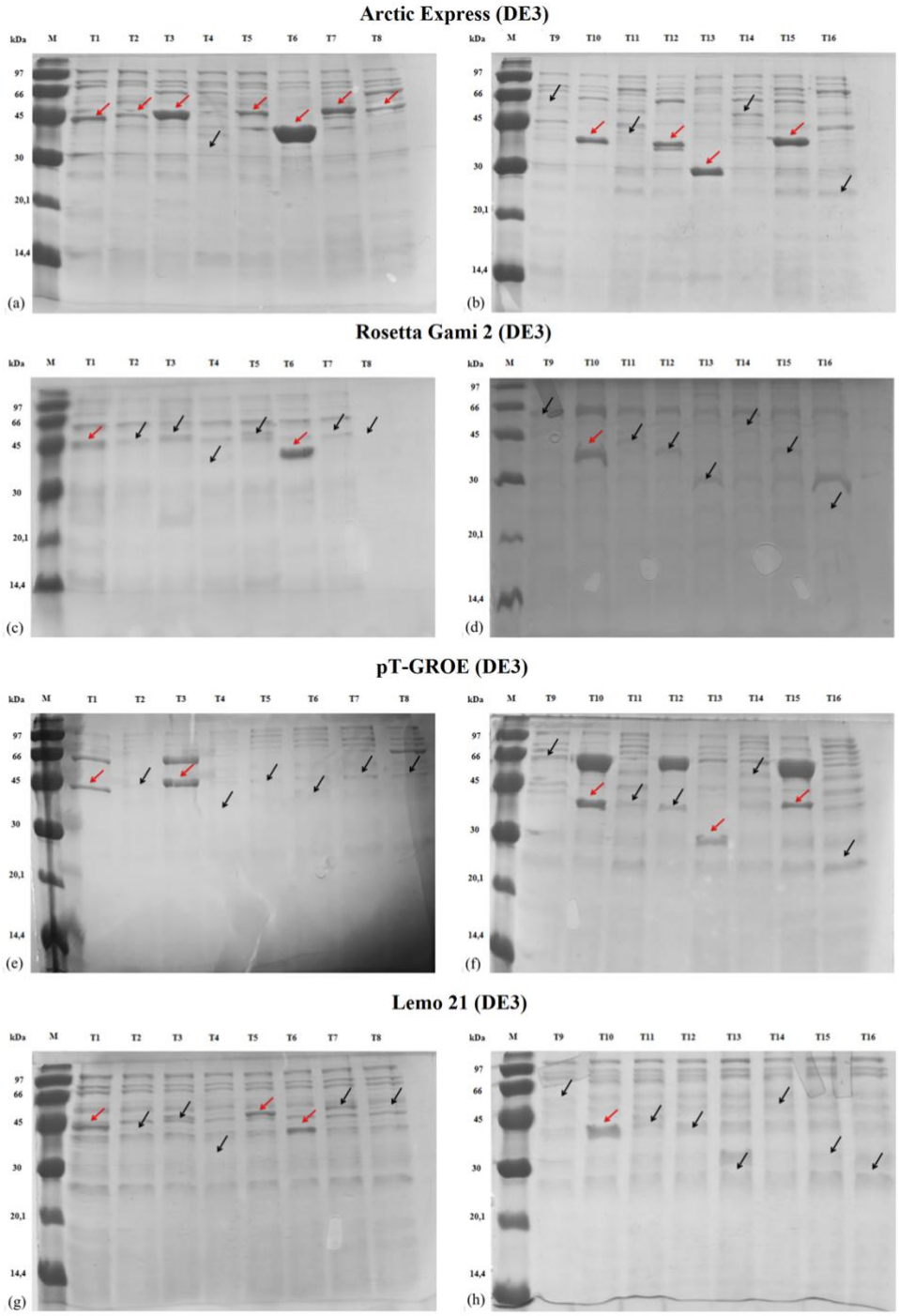

Figure 1. SDS-PAGE (15%) analysis of soluble protein recovery in the four strains used in this work. The red arrows indicate the recovered protein with the correct molecular weight. The black arrows indicate the molecular weight of the proteins that were not recovered or only minimally recovered.

The targets **T3**, **T4**, **T5**, **T6,** and **T10** are five enzymes of the same pathway in *M. tuberculosis* and *Plasmodium* species. Out of these five targets, the Arctic strain recovered three of them in the soluble form and a fourth target (**T5**) to a lesser extent, while pT-GroE recovered two of them with a very good expression pattern (**T3** and **T10**). Lemo21 strain recovered **T5**, **T6**, and **T10**, and Rosetta Gami 2 recovered targets **T6** and **T10**. The targets **T8** and **T9** were recovered in very low yields in the pT-GroE strain. The Arctic Express strain also recovered slow amounts of **T8**, while Rosetta Gami 2 recovered small amounts of **T9**.

The targets **T11** and **T14** are two glycosyltransferases, which are targets known to be insoluble in many cases. **T11** was also tested for its expression in two truncated forms (**T12** and **T13**) and also in the full-length protein in the pET-NESG vector (**T16**). A truncated form of **T14** was also tested in the pETTRXA-1a/LIC vector and is listed as **T15**. The Lemo21 strain was not able to recover any of the targets, while Rosetta Gami recovered only **T13** to a very small extent. The pT-GroE strain recovered two of the truncated forms for these targets (**T13** and **T15**) and, to a lesser extent, the third truncated form (**T12**). Finally, the Arctic strain recovered the same three truncated forms with very good contrast to the background proteins observed in the SDS-PAGE.

After screening the strains available in our group, we found, in general, that 9 out of the 16 targets, or 56% were recovered in reasonable amounts. In at least five cases, the recovery was observed in multiple strains, while leastwise in three cases, a single strain was found to be more efficient in recovering a target in the soluble form (**T7**, **T13**, and **T14**). In the other five cases, including the two full-length glycosyltransferases, none of the strains used here was found to recover the target in the soluble form.

To be able to test several different targets in a batch experiment, we missed some optimization procedures that could enhance our recovering rate. For example, the Lemo21 strain can fine-tune the expression level of toxic protein by the titratable regulation of the T7 lysozyme (T7L) expression using L-rhamnose [10,11].

pT-GroE, a strain that co-expresses a molecular chaperone to assist the proper protein folding, was found to result in a good recovering rate as well as a good contrast from the target protein to the basal protein expression, making it very easy to follow the target protein being expressed. Similar to pT-GroE, the Arctic Express strain combines

low temperature-adapted chaperones Cpn10 and Cpn60, similar to *E. coli* GroEL and GroES, to result in chaperon assisted and low-temperature protein expression [12]. This strategy was successful in several targets tested here and was also the strain with a higher recovering rate.

The recovering rate in any structural biology pipeline will always be highly dependent on the targets chosen in the project. In this sense, we have no aim to claim that a given commercial strain is better or worse than each other. Rather than that, we show that it might be worth having a suitable and scalable assay to screen different host strains for the expression of soluble protein in difficult cases. We easily foresee a scenario where proteins that were not recovered in the soluble fraction after a '*standard'* protocol can be re-assayed for a '*salvage'* pipeline, where different strains are tested for the expression. The experimental cost of this includes transforming the existing vector in different cell strains and a protocol for expression and purification that is scalable and may achieve an HT format. Cases that cannot be recovered in this salvage pipeline are still suitable for changing the construction or even going for a eukaryote cell as the host in heterologous expression. In these cases, the experimental costs are slightly higher and involve going back to construct and primer design, gene cloning, and testing for expression again.

4. Conclusion

In conclusion, here we showed that nine out of 16 'difficult' targets could be recovered in the soluble fraction after testing a set of different *E. coli* strains. The assay is easy to implement and scalable for a different number of genes or strains. We found that the screening may be an affordable salvage pipeline for the recovery of target proteins after a typical pipeline for protein expression and purification in structural biology efforts.

Acknowledgments

The authors thank to the members of the Molecular Biotecnology Group @ IFSC/USP and, in particular, for our technical staff: Maria Auxiliadora Santos, João

Possatto and Josimar Sartori. The authors also thank the financial support from the Fundação de Apoio À Pesquisa do Estado de São Paulo (FAPESP) through grants, 2017/24901-8, 2018/21213-6, 2019/20219-3 and 2020/03983-9 and from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through grants 303165/2018-9, 485950/2013-8 and 476606/2010-1. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] C.R. BEDDELL, P.J. GOODFORD, F.E. NORRINGTON, S. WILKINSON, R. WOOTTON, COMPOUNDS DESIGNED TO FIT A SITE OF KNOWN STRUCTURE IN HUMAN HAEMOGLOBIN, Br. J. Pharmacol. 57 (1976) 201–209. doi:10.1111/j.1476-5381.1976.tb07468.x.

[2] G.L. Rosano, E.A. Ceccarelli, Microbial Cell Factories Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain, (2009). doi:10.1186/1475-2859-8-41.

[3] J. Kaur, A. Kumar, J. Kaur, Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements, Int. J. Biol. Macromol. 106 (2018) 803–822. doi:10.1016/j.ijbiomac.2017.08.080.

[4] M.T. Sonoda, A.S. Godoy, V.O.A. Pellegrini, M.A.S. Kadowaki, A.S. Nascimento, I. Polikarpov, Structure and dynamics of Trichoderma harzianum Cel7B suggest molecular architecture adaptations required for a wide spectrum of activities on plant cell wall polysaccharides, Biochim. Biophys. Acta - Gen. Subj. 1863 (2019) 1015–1026. doi:10.1016/j.bbagen.2019.03.013.

[5] C.M. Camilo, I. Polikarpov, High-throughput cloning, expression and purification of glycoside hydrolases using Ligation-Independent Cloning (LIC)., Protein Expr. Purif. 99 (2014) 35–42. doi:10.1016/j.pep.2014.03.008.

[6] L. Stols, M. Gu, L. Dieckman, R. Raffen, F.R. Collart, M.I. Donnelly, A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site, Protein Expr. Purif. 25 (2002) 8–15. doi:10.1006/prep.2001.1603.

[7] W.H. Eschenfeldt, N. Maltseva, L. Stols, M.I. Donnelly, M. Gu, B. Nocek, K. Tan, Y. Kim, A. Joachimiak, Cleavable C-terminal His-tag vectors for structure determination, J. Struct. Funct. Genomics. 11 (2010) 31–39. doi:10.1007/s10969-010-9082-y.

[8] W.H. Eschenfeldt, L. Stols, C.S. Millard, A. Joachimiak, M.I. Donnelly, A Family of LIC Vectors for High-Throughput Cloning and Purification of Proteins 1, (n.d.). doi:10.1007/978-1-59745-196-3_7.

[9] F.W. Studier, Protein production by auto-induction in high-density shaking cultures, (2005). doi:10.1016/j.pep.2005.01.016.

[10] S. Schlegel, J. Löfblom, C. Lee, A. Hjelm, M. Klepsch, M. Strous, D. Drew, D.J. Slotboom, J.W. De Gier, Optimizing membrane protein overexpression in the Escherichia coli strain Lemo21(DE3), J. Mol. Biol. 423 (2012) 648–659. doi:10.1016/j.jmb.2012.07.019.

[11] S. Schlegel, E. Rujas, A.J. Ytterberg, R.A. Zubarev, J. Luirink, J.W. de Gier, Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels, Microb. Cell Fact. 12 (2013) 24. doi:10.1186/1475-2859-12-24.

[12] M. Ferrer, T.N. Chernikova, M.M. Yakimov, P.N. Golyshin, K.N. Timmis, Chaperonins govern growth of Escherichia coli at low temperatures, Nat. Biotechnol. 21 (2003) 1267–1267. doi:10.1038/nbt1103-1266b.

[13] ProtParam References, (n.d.). https://web.expasy.org/protparam/protpar-ref.html (accessed February 26, 2021).

Table 1. Target genes used for protein expression. The isoelectric point and molecular mass were computed with the ProtParam [13] server. TBP = thiamine biosynthesis pathway. PBP = pyridoxal 5'-phosphate biosynthesis pathway. GT= glycosyltransferase.

Target	Pathway	Organism	N_{res}[a]	N_{Cys}[a]	pI[a]	MW[b] (kDa)
T1	TBP	*E. faecalis*	276	1	5.47	43.3
T2	TBP	*S. aureus*	276	2	5.78	44.2
T3	PBP	*M. tuberculosis*	299	1	5.24	45.4
T4	PBP	*M. tuberculosis*	199	2	5.33	35.2
T5	PBP	*P. vivax*	302	7	6.62	47.2
T6	PBP	*P. vivax*	219	7	6.54	38.5
T7	TBP	*P. falciparum*	310	16	6.18	49.0
T8	TBP	*P. falciparum*	302	11	8.88	47.9
T9	TBP	*P. falciparum*	400	13	8.89	60.7
T10	PBP	*P. falciparum*	219	9	6.43	38.6
T11	GT	*E. faecalis*	241	2	6.23	40.9
T12	GT	*E. faecalis*	210	2	6.22	37.7
T13	GT	*E. faecalis*	130	1	4.39	28.4
T14	GT	*E. faecalis*	325	0	9.08	51.2
T15	GT	*E. faecalis*	220	0	5.59	39.1
T16	GT	*E. faecalis*	241	2	6.23	26.9

[a] Applies to the target sequence, without considering the fusion protein sequence.

[b] Molecular mass computed with the fusion tag.

Table 2. Summary of the soluble protein recovery using different *E. coli* strains culture medium. The expression levels, as shown in Figure 1, were quantified in a scale from (+) to (++++). Here, (-) indicates no expression.

	***E. coli* strains**			
Target	Arctic Express	Rosetta Gami 2	pT-GroE	Lemo 21
T1	++	++	++	++
T2	+	+	-	-
T3	+++	+	+++	+
T4	-	-	-	-
T5	++	+	-	++
T6	++++	++	-	++
T7	++	+	+	+
T8	+	-	+	-
T9	-	+	+	-
T10	++	++	++	++
T11	-	-	+	-
T12	++	+	+	-
T13	+++	+	++	-
T14	+	-	+	-
T15	+++	-	++	-
T16	-	-	+	-

* The recovery yield levels are ranked from no expression (-) to high yield (++++).