

Chapter XXX.

Overview of a high throughput pipeline for streamlining the production of recombinant proteins

Running title: a pipeline for protein production

Joanne E. Nettleship, Heather Rada and Raymond J. Owens

Research Complex at Harwell, Rutherford Appleton Laboratory Harwell Oxford, UK and Division of Structural Biology, Henry Wellcome Building for Genomic Medicine, University of Oxford, Roosevelt Drive, Oxford, UK

Corresponding author : Raymond J Owens

Abstract: Production of high quality protein is an essential step for both structural and functional studies. Throughput has increased in the past decade by the use of streamlined workflows with standard operating procedures and automation. In this chapter, we describe the Oxford Protein Production Facility (OPPF) pipeline for protein production, from conception, through vector construction, to expression and purification. Results from projects run in the OPPF demonstrate the value of using parallel expression screening of intracellular proteins in both *E. coli* and insect cells. Transient expression in Human Embryonic Kidney (HEK) cells is used exclusively for production of secreted glycoproteins. Protein purification and quality assessment are independent of the expression system and enable sample preparation to be simplified and streamlined.

Key words : recombinant protein production, high throughput, *E. coli*, insect cells, HEK cells

Overview of a high throughput pipeline for streamlining the production of recombinant proteins

Abstract

Production of high quality protein is an essential step for both structural and functional studies. Throughput has increased in the past decade by the use of streamlined workflows with standard operating procedures and automation. In this chapter, we describe the Oxford Protein Production Facility (OPPF) pipeline for protein production, from conception, through vector construction, to expression and purification. Results from projects run in the OPPF demonstrate the value of using parallel expression screening of intracellular proteins in both *E. coli* and insect cells. Transient expression in HEK cells is used exclusively for production of secreted glycoproteins. Protein purification and quality assessment are independent of the expression system and enable sample preparation to be simplified and streamlined.

Key words : recombinant protein production, high throughput, *E. coli*, insect cells, HEK cells

1. Introduction

The production of high quality recombinant proteins is a critical step in many fields of protein science but especially structural biology where sample quality is crucial for downstream applications including crystallization, cryo-electron microscopy and nuclear magnetic resonance measurements. Over the past decade, the production of proteins for structural studies has been streamlined by the introduction of standard operating procedures and the use of laboratory automation to enhance throughput. The different stages in design, expression and purification of recombinant proteins at laboratory scale have become integrated into a single workflow. The pipeline developed by the Oxford Protein Production Facility (OPPF) (Fig. 1) is an example of this approach and comprises three stages (1) construct design (2) expression screening and (3) sample preparation. Each of these steps are considered in the sections that follow.

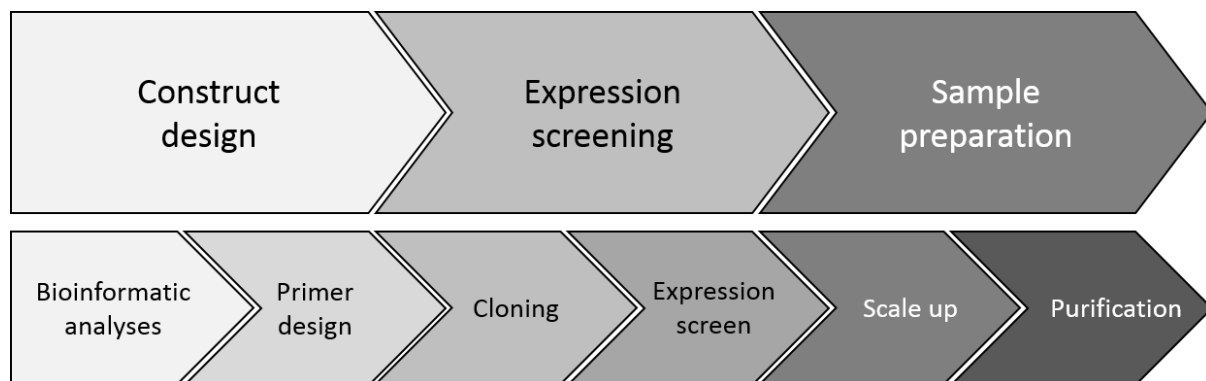


Figure 1

2. Construct Design

Designing the construct(s) that will encode the target protein is probably one of the most critical stages in the process of producing a recombinant protein. Therefore, at the outset, it is important to define the intended use of the protein, for example activity screening for which only the catalytic domain is required or the full length open reading frame to determine the overall architecture of the protein. The approach taken in the OPPF typically involves making both full length and truncated constructs in order to maximise the output of the experiment.

2.1 Bioinformatic analyses

Information about the target protein(s) is gathered from the scientific literature and by various bioinformatics analyses as part of the design process. Resources such as Uniprot and the Protein Databank (PDB) are used as entry points to prior knowledge. In addition, selected online resources are used to gather information, for example disorder prediction using the RONN algorithm (Fig. 2) [1] and to generate 3-D homology models using the Phyre2 server [2].

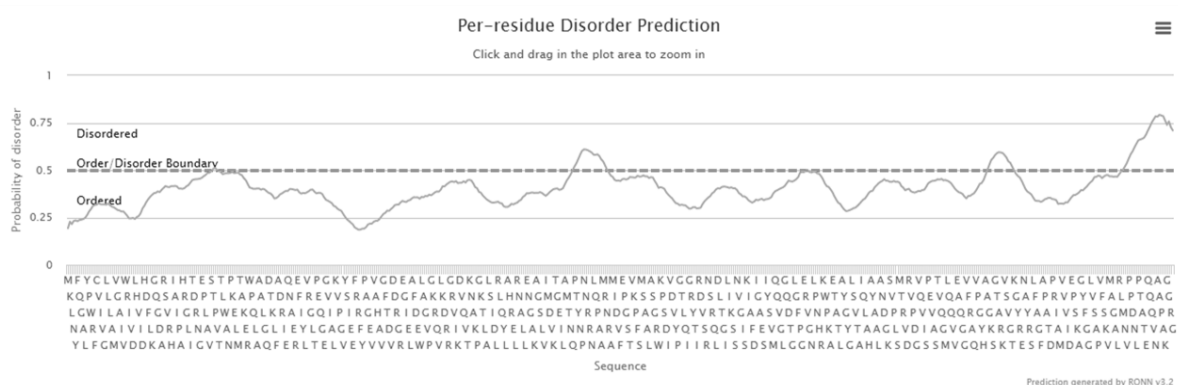


Figure 2

A list of the web-based bioinformatic resources routinely used in the OPPF is given in Table 1.

2.2 Primer design

The information compiled from both the scientific literature and bioinformatics analyses is used to inform the choice of start and stop positions of expression constructs and hence the locations of the forward and reverse primer sequences that will be used for PCR amplification and cloning. An in-house MySQL database has been developed, called OPTIC [3] for storing target sequences. On entry into OPTIC each sequence acquires a unique identifier (OPTIC number). A primer design tool linked to the OPTIC database enables PCR primers to be designed automatically but also manually adjusted as required. The OPTIC database contains a table of all the primer extensions required for ligation independent cloning and these are automatically added to the gene-specific forward and reverse primer sequences by selecting the vector that will be used for cloning and expression screening (see below). When a construct is generated it is automatically given a unique OPPF number and the PCR primer sequences stored alongside the sequence that will be amplified. Typically more than one construct is designed for each target protein such that there is a one to many relationship between target (OPTIC number) and constructs (OPPF numbers). The list of OPPF numbers and associated OPTIC identifiers can be exported from the OPTIC database into a construct design spreadsheet, with each OPPF construct corresponding to a position in a 96-well plate (A1-H12). An order form for PCR primers can be created either manually by “cutting and pasting” from the record in OPTIC or automatically using a protocol written in the Protein Information Management System, PiMS [4].

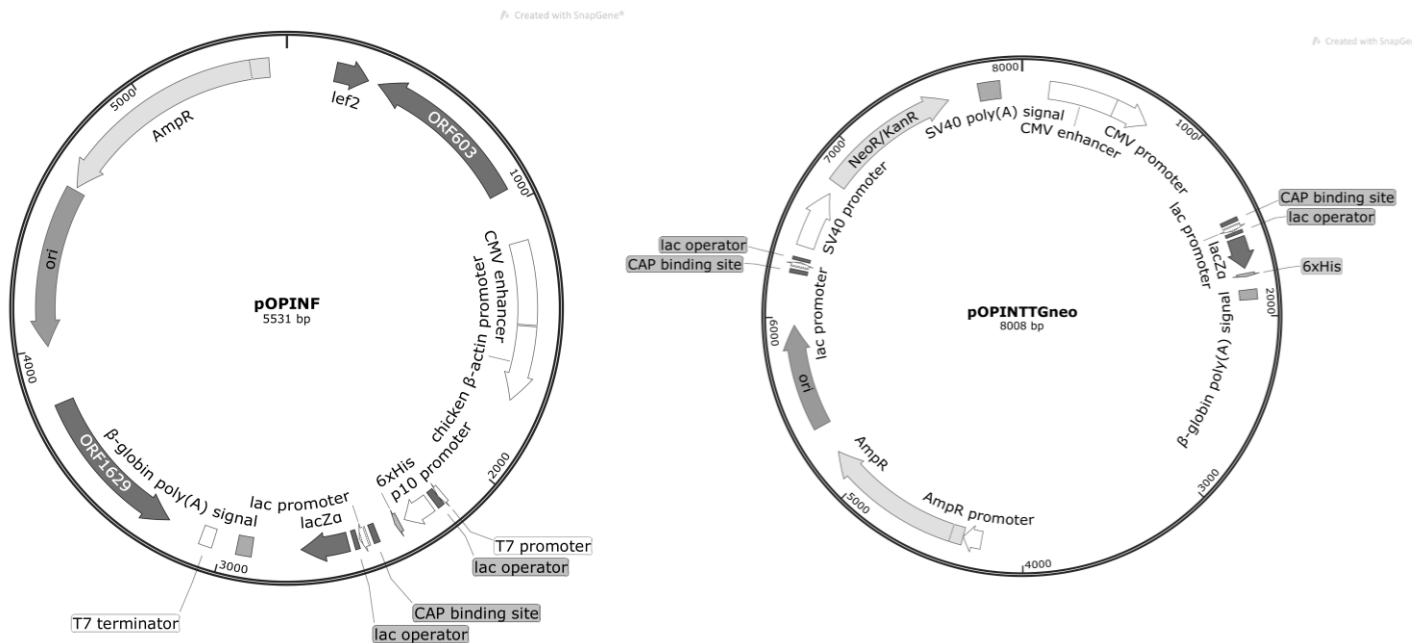
3. Expression screening

For all cloning and expression experiments, constructs are organised in a 96 well SBS plate format as defined by the construct design spreadsheet. Each experiment typically comprises between 24 and 96 expression vectors corresponding to one or more targets. The multiple constructs for each target may consist of different domains or sub-domains and/or different fusion tags added to facilitate protein

solubility and detection/purification. Subsequently, the different vector configurations can be tested for expression in different hosts (*E. coli*, insect and mammalian cells). In this way information is obtained about the optimal construct and expression host for subsequent production of the protein target.

3.1 Vector construction

Multiple expression vectors are constructed in parallel by ligation independent cloning using the Infusion™ polymerase that catalyses the precise joining of DNA fragments (e.g. PCR-generated inserts, synthetic genes and linearized vectors) by recognizing 15 bp overlaps at their ends [5]. The primer design tool automatically adds the appropriate 15 bp extension that is required for cloning into a particular vector. A suite of vectors (the pOPIN vectors) based on the pTriEx 1.1 have been constructed such that the same 15 bp annealing sequences can be used for cloning. Hence the same PCR product can be used to construct a whole series of pOPIN vectors that in turn can be tested in more than one host [5-8]. The pTriEx plasmid backbone incorporates promoter elements for expression in *E. coli*, mammalian cells and insect cells using baculoviruses (Fig. 3A). Therefore the same vector, generated by a single cloning step can be used for screening in different host cells. For expression in insect cells, homologous recombination within the insect cell is used to construct baculoviruses in a single step. The pOPIN transfer vector is co-transfected with a genetically disabled baculovirus propagated as a bacmid and the recombinant virus subsequently amplified [9, 10]. All the pOPIN vectors include a hexa/octahistidine tag to facilitate detection and purification of expressed proteins. A subset of pOPIN vectors have been constructed for production of secreted eukaryotic proteins that have a resident signal sequence [11] and are based on the pTT vector backbone [12] (Fig. 3B). In addition, a number of fusion protein tags have been added that can aid solubility and/or expression levels (Fig. 4).



Figures 3A & 3B

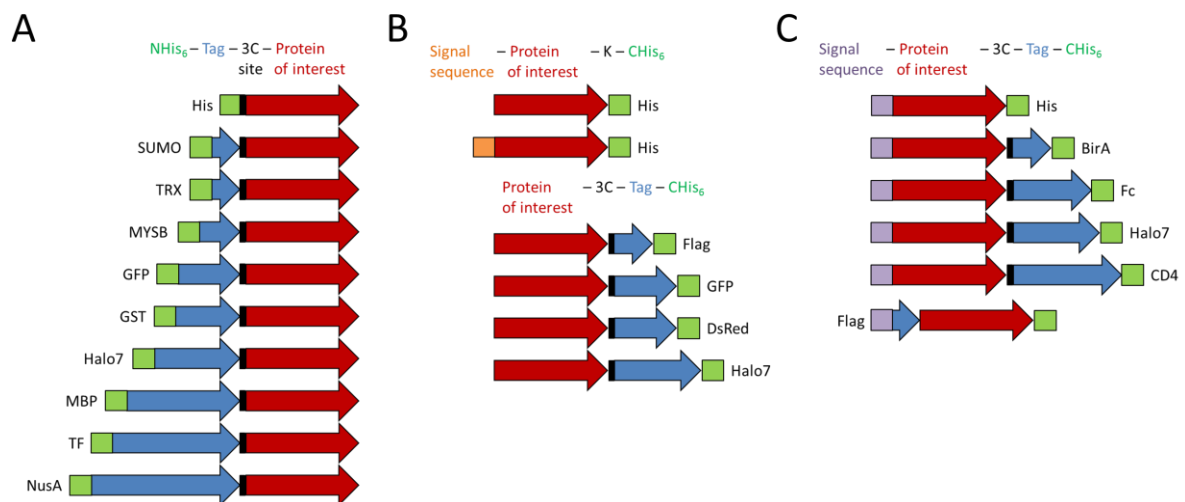


Figure 4

3. 2. Screening for the expression of intracellular proteins

Expression screening of prokaryotic proteins is carried out exclusively in *E. coli*. However, for the production of eukaryotic proteins, experience has shown that testing for expression in both *E. coli* and insect cells significantly increases the number of expressed targets. In both cases, expression screening is automated and performed in 96-well format. The assay consists of pelleting 1 ml of

culture, lysing the cells and performing small scale nickel affinity capture using magnetic beads [5] (Fig. 5). The resulting purified protein is then analysed by SDS-PAGE and bands compared to a GFP (green fluorescent protein) positive control. Expression screening in *E. coli* and insect cells can take place in parallel, although the insect cell screen takes longer to perform due to raising the baculovirus (Fig. 5).

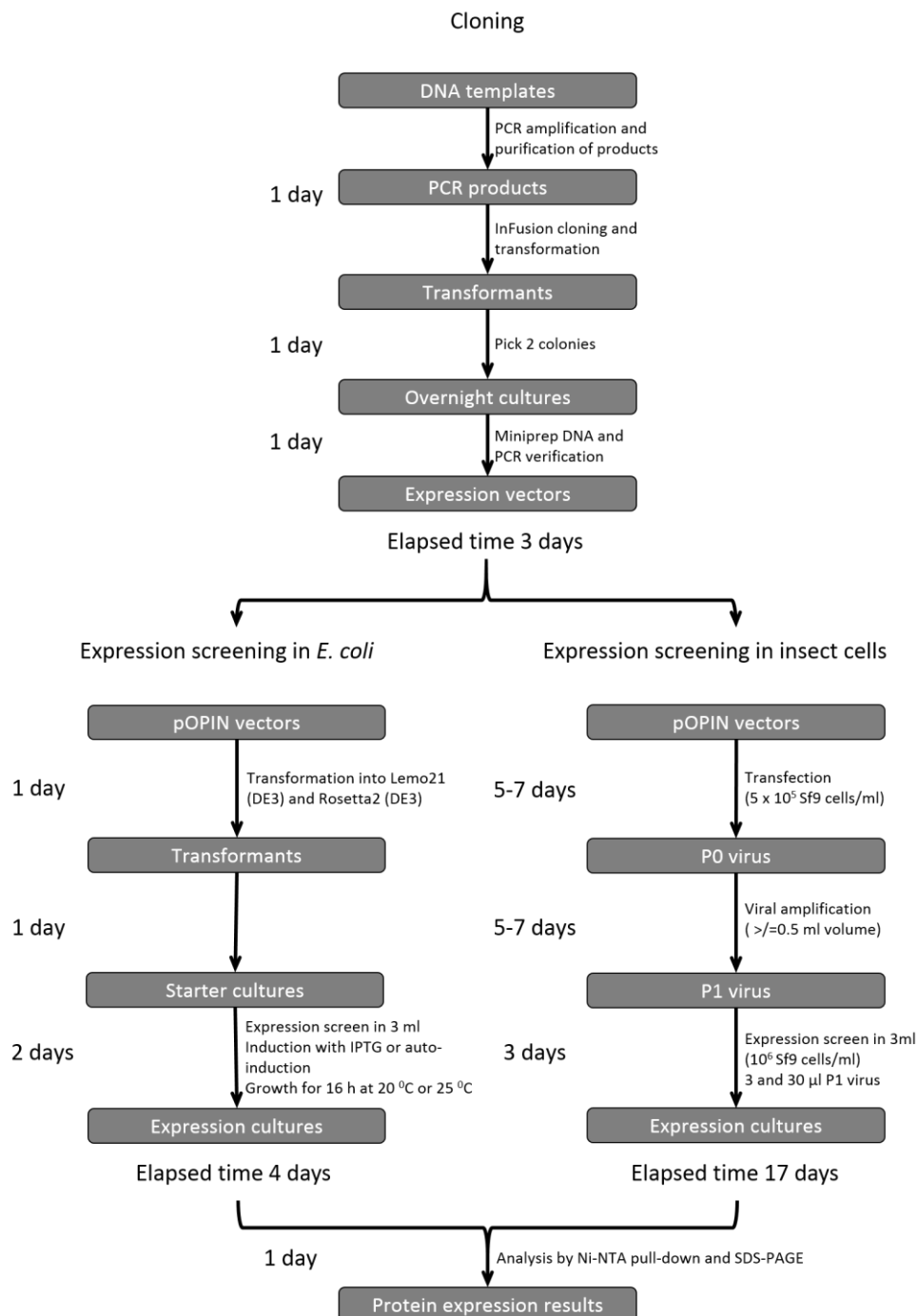
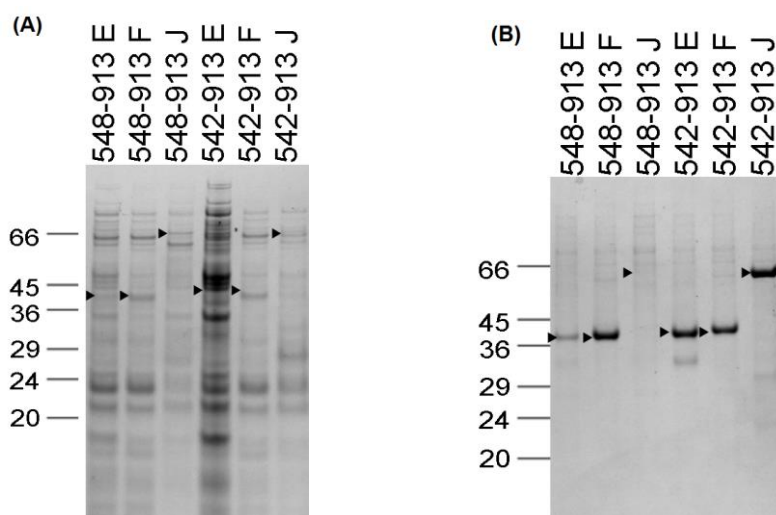


Figure 5

An example of the results obtained from using both *E. coli* and insect cells in expression trials run in parallel is shown in Figure 6. Here, no detectable expression of Discoidin Domain Receptor Tyrosine Kinase 1 (DDR1) was seen in *E. coli* (Fig. 6A) whereas protein of the expected size was expressed in insect cells infected with recombinant DDR1 kinase baculoviruses (Fig. 6B).



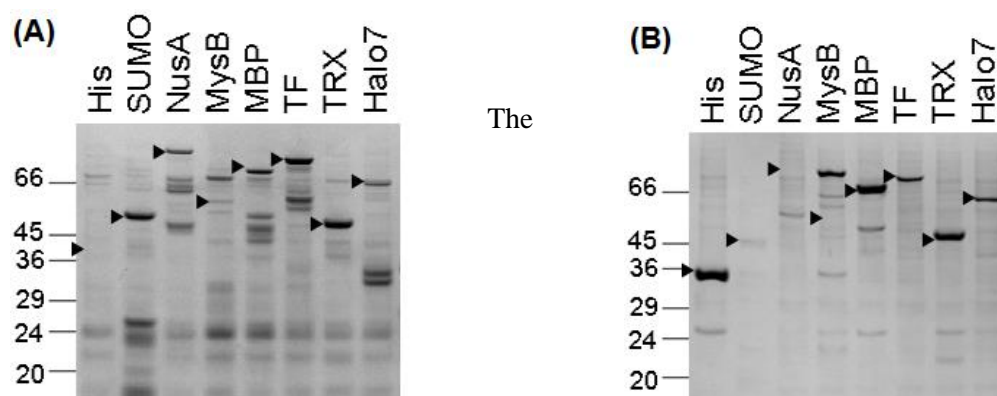
Figures 6A and 6B

The results for screening soluble protein expression in both *E. coli* and insect cells of 427 constructs corresponding to 135 different eukaryotic proteins are shown in Table 2. Constructs contained either an N or C-terminal hexhistidine tag. A total of 207 constructs representing 65 proteins (unique OPTIC numbers) gave expression in either *E. coli*, insect cells or both. However, of these only 58 (28 %) would have been obtained using *E. coli* alone. Whereas insect cells alone accounted for 183 (88 %) of the expression hits. For eukaryotic proteins, using insect cells for expression is clearly beneficial. Screening using both *E. coli* and insect cells in parallel will maximise the number of expression positive constructs obtained in the shortest possible time.

Fusion of a target protein to a highly soluble carrier protein can enhance the expression of eukaryotic proteins in *E. coli*. Figure 7A shows the full length construct of a phospholipase C family protein expressed in *E. coli* with 8 different N-terminal tags. In terms of expression in *E. coli*, it can be seen that the His-tag alone gave no detectable expression for this protein, whereas addition of solubility tags led to detectable expression of the fusion protein (Fig. 7A). However, one of the drawbacks of using solubility tags is that truncated products are often observed, for instance in the lane with the

Halo7 tagged protein, bands can be seen around 35 kDa which correspond to the Halo7 tag alone.

When the same constructs were expressed in insect cells not only the fusion proteins, but also the His-tagged version were expressed (Fig. 7B). This construct was scaled to 1 L in insect cells and gave 1.4 mg of purified protein.



Figures 7A and 7B

The conclusion from screening expression of 135 proteins (unique OPTIC numbers used in this dataset) is that changing the expression system is better than trying to produce *E. coli* fusion proteins.

3. 3. Screening for the expression of secreted glycoproteins proteins

Production of secreted (glyco)proteins forms a major part of the project portfolio of the OPPF for which transient expression in HEK 293 cells is routinely used [13]. As with intracellular proteins, the same principal of evaluating multiple constructs in parallel is applied to secreted proteins. Cells are grown in either 24- or 96-well tissue culture plates and expression of secreted proteins detected by western blotting and/or ELISA of culture supernatants 72 hours post-transfection. For projects requiring multiple rounds of screening, the transient transfection process has been automated [10].

In most cases, proteins are expressed using the RTPT μ signal peptide [11] from the pOPINTTGneo vector, but native signal peptides have also been tested using the pOPINeNeo vector. With secreted proteins, the addition of a C-terminal CD4 fusion protein [14] can improve relative expression levels as in the case shown in Figure 8. Therefore, this tag is routinely tested alongside the C-terminal His tag. In this example, the native signal sequence did not lead to detectable secreted protein in contrast

to the RTPT μ signal peptide (Fig. 8). Where tested, we have found that RTPT μ has invariably given the same or better secretion than the native signal peptide.

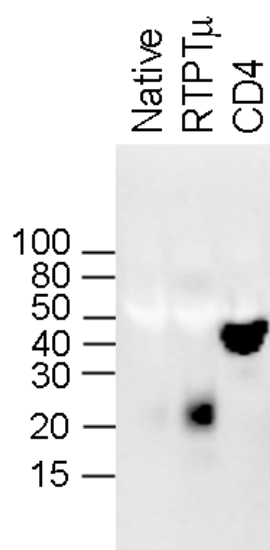


Figure 8

4. Sample preparation

In most cases the results from small-scale expression screening translate to the outcomes at larger scale. Overall screening enables effort to be prioritised so that resources (time and money) are only invested in those constructs that will give sufficient yields for downstream applications.

4.1 Scale up and purification

The results from expression screening inform the choice both of construct and production system for sample preparation. Simple fed batch cultures are grown for *E. coli*, insect and mammalian cells using commercially available media in 1- 2.5 L volumes.

Expression pre-screening means that a simple two-step purification method is sufficient to prepare samples of sufficient purity for structural studies (Fig. 9). For intracellular proteins, the first step is a nickel column followed automatically by a size exclusion column using an ÄktaXpress purification system (GE Healthcare). If a cleavable tag has been used, this is then cut using 3C protease before secondary purification (Fig. 9A). Products secreted from mammalian cells undergo the same initial

purification steps of nickel column followed automatically by size exclusion using the ÄktaXpress programme published by Nettleship *et al.* (Fig. 9B) [15]. This programme allows large volumes of media to be loaded onto the nickel column using a cycle of loading 200 ml media followed by 50 ml of wash buffer before continuing with elution from the nickel column and size exclusion chromatography. The method reduces pressure build-up due to viscous components of the media alongside reducing non-specific binding to the nickel column. Non-specific binding is also reduced by the addition of 2 mM NiCl₂ to serum-containing media before loading onto the nickel column. If required, N-glycosylation of the product can be simplified by the addition of kifunensine to the cell media during scale-up of expression [16]. Kifunensine inhibits α -mannosidase I and leaves high-mannose N-glycans on the expressed glycoproteins which can then be removed after purification by treatment with Endoglycosidase F1 (EndoF1) followed by a second gel filtration step to remove the EndoF1 from the sample.

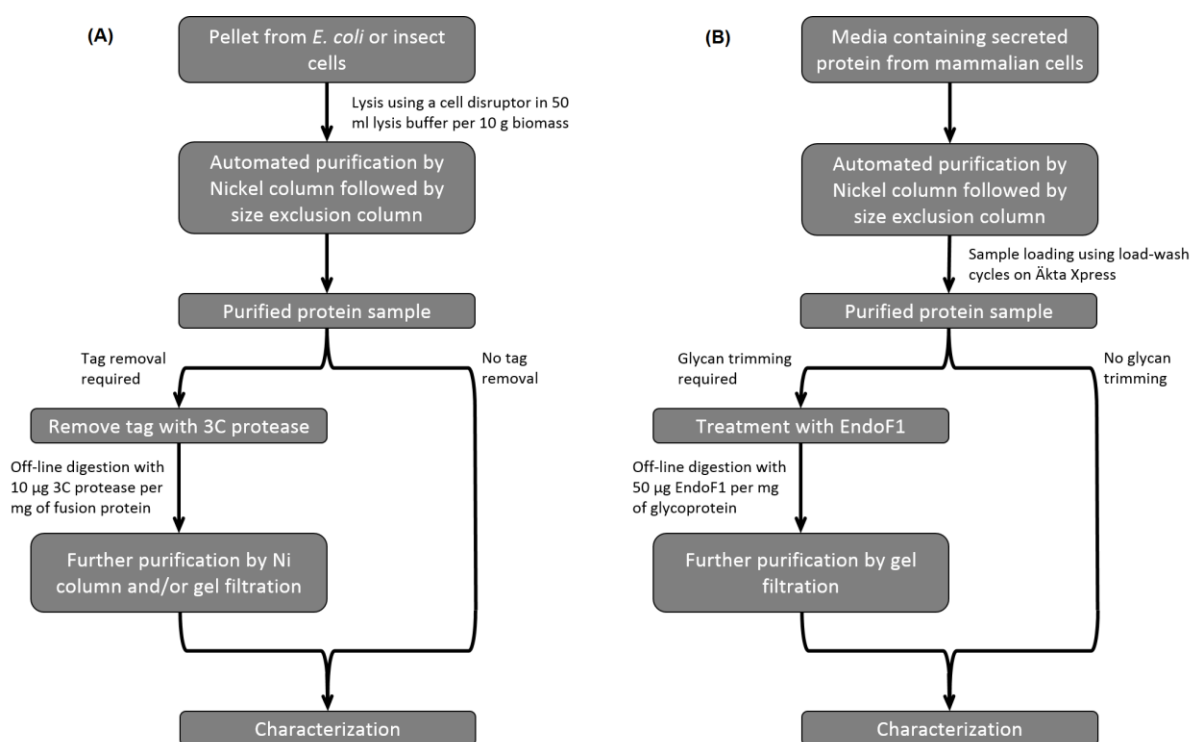


Figure 9

A standard set of buffers is used for the column chromatography steps which simplifies the process and facilitates running several protein purifications in parallel. The buffers used for nickel affinity

chromatography contain 50 mM Tris, pH 7.5 and 500 mM NaCl with 30 mM or 500 mM imidazole, pH 7.5 for the wash and elution buffers respectively. The size exclusion buffer is 20 mM Tris, pH 7.5 and 200 mM NaCl with 1 mM TCEP (Tris(2-carboxyethyl)phosphine hydrochloride) if a reducing agent is required. Examples of size exclusion profiles from one week of protein purification using the standard protocols and buffers are shown in Fig. 10. The majority of the gel filtration profiles (HiLoad 16/600 Superdex 200 or Superdex 75, GE Healthcare) show symmetrical peaks corresponding to a relatively monodisperse product, and further analysis by SDS-PAGE showed that the proteins were > 95 % pure and hence suitable for biophysical characterisation, activity assay and/or crystallization. The profile for protein OPPF 17541 shows two peaks which upon further analysis (not shown here) related to monomeric and dimeric forms of the protein. OPPF 16681 can be seen to give many peaks in the size exclusion profile and therefore needs further purification.

Using the standard buffers, 76 % of proteins entering the sample preparation stage in the OPPF have been successfully purified. This success rate is independent of whether the protein is intracellular or secreted and of which host cell was used for production. Of the samples that failed production, unexpectedly low initial expression level on scale-up, precipitation upon tag removal or aggregation of the protein were among the most common problems encountered. Low yield can be addressed by significantly increasing culture volumes for scale-up to produce more biomass and protein solubility by optimising buffer conditions (see below).

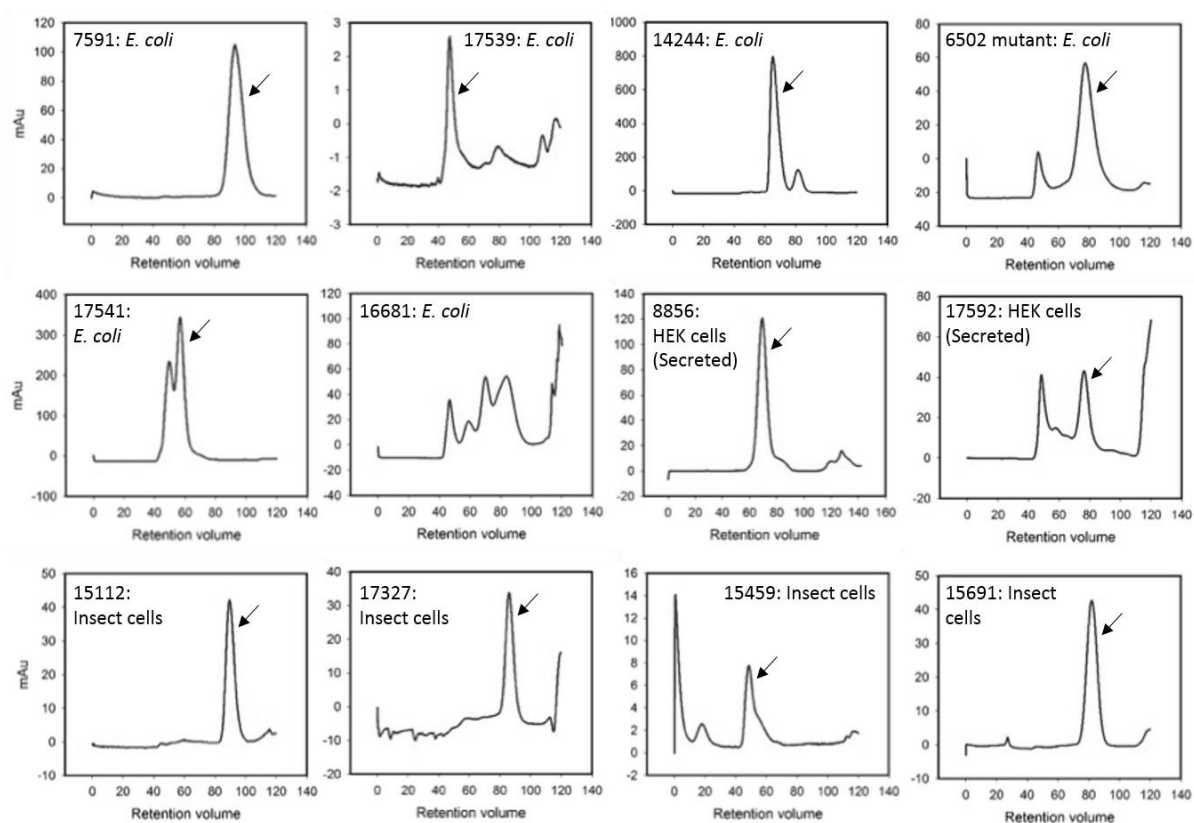


Figure 10

4.2 Biophysical characterisation

At the OPPF, all purified proteins are characterised by intact protein mass spectrometry as a quality assurance step [17]. This allows the measured mass to be compared with the expected mass and any post-translational modifications to be assessed. Figure 11 shows intact protein mass spectrometric analyses of a number of purified proteins produced from different host cells. Mass spectrometry can be used to measure modifications, for example OPPF 2145 where the measured mass relates to labelling of the three methionines with selenomethionine (Fig. 11). Glycoproteins are treated with PNGase F prior to intact protein mass spectrometry to remove the N-glycans. This alters the asparagine at the N-glycosylation site to an aspartic acid resulting in a +1 Da increase in mass. OPPF 19763 has a +1 Da shift in mass relating to one N-glycosylation site and OPPF 7040 has a +2 Da shift showing that it contains two occupied N-glycosylation sites (Fig. 11). The mass spectrum for OPPF 20371 + 18732 shows one peak corresponding to a non-reduced Fab fragment where the heavy and light chains are linked by a cysteine bridge. This was produced by co-transfection in mammalian cells

[18]. In the case of OPPF 3325, although purified protein could be seen by SDS-PAGE, the protein failed quality assurance by mass spectrometry (Fig. 11).

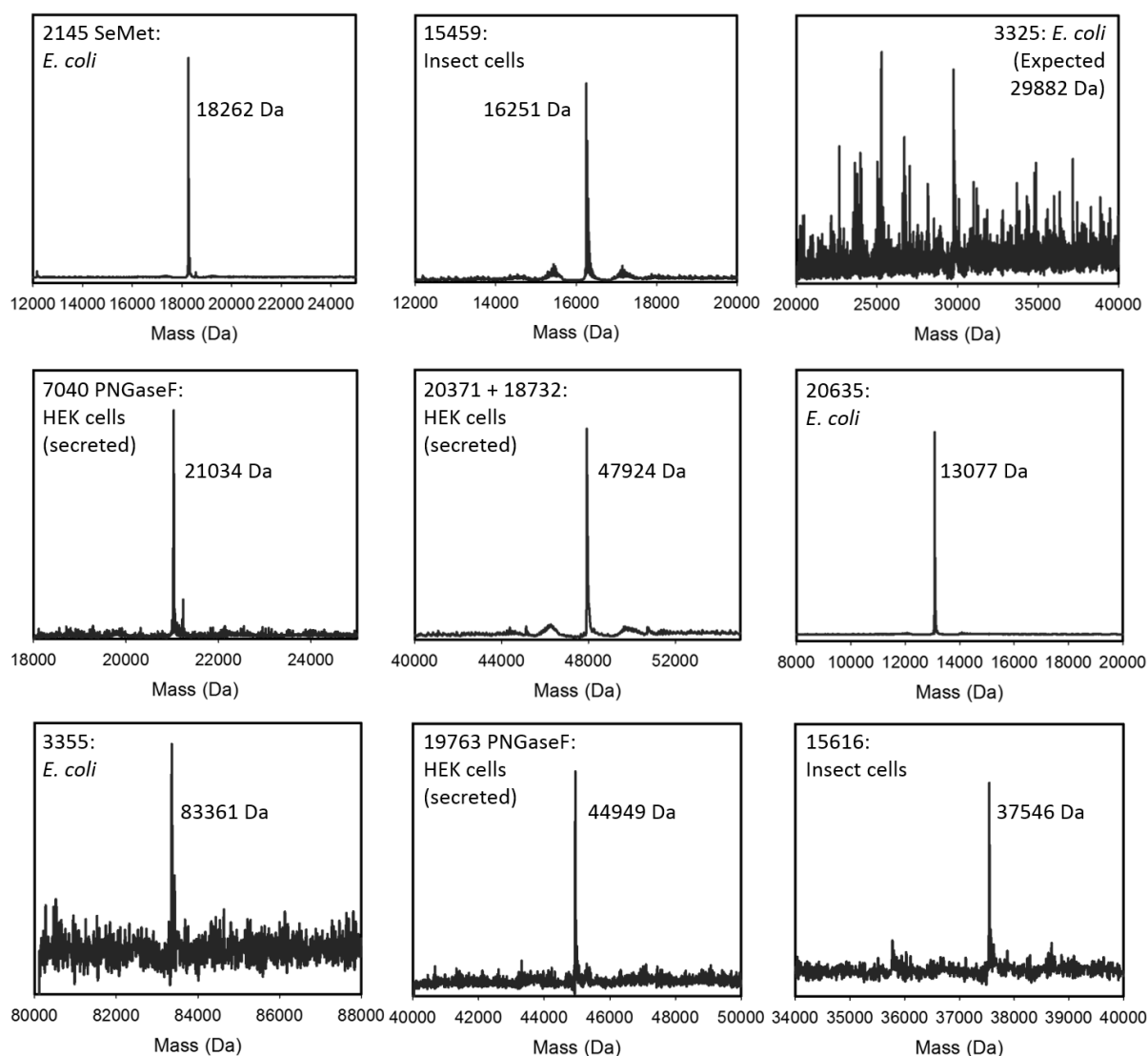


Figure 11

The second sample quality measure that is routinely used is the thermal shift assay [17]. Although not applicable to all proteins, this assay is a fast and convenient method for indirectly assessing protein folding. Further, by analysing the thermal stability of a sample in different buffer conditions, formulation of the sample can be optimised.

Figure 12 shows results from a thermal shift assay for a human histone acetyltransferase. Here it can be seen that there is a shift to a higher melting temperature, and therefore increased protein stability,

at pH 6.5 (Grey dashed line; $T_m = 39.9\text{ }^{\circ}\text{C}$) rather than the standard pH of 7.5 (Black solid line; $T_m = 38.6\text{ }^{\circ}\text{C}$). This information is used to determine the optimum pH for purification and storage of the protein.

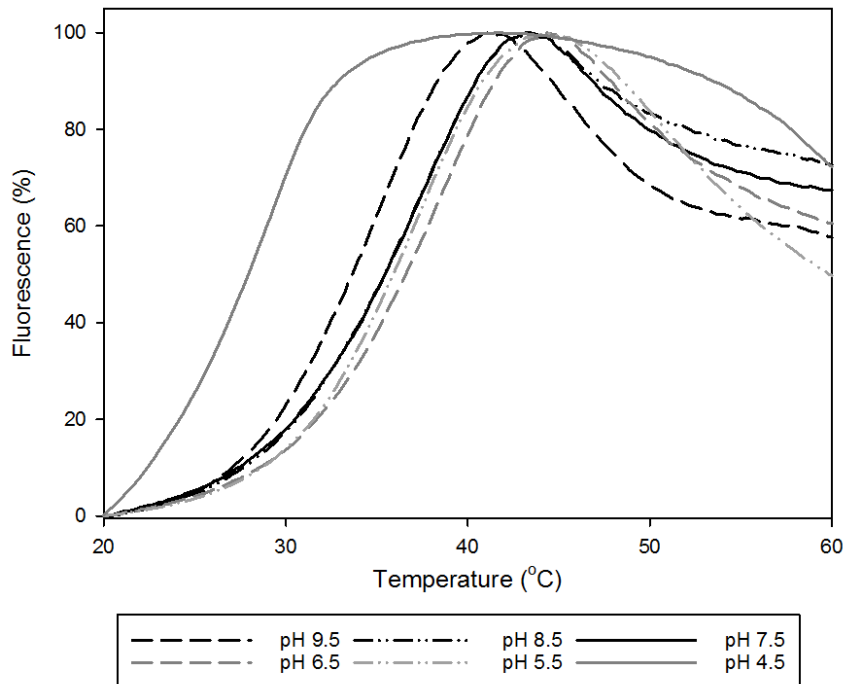


Figure 12

5.

Summary

The OPPF pipeline has developed in order to streamline protein production for structural and functional studies. A number of factors, outlined below, make this possible:

- 1) Design of multiple constructs at the start of a project.
- 2) Ligation independent cloning in 96-well format.
- 3) Preferential use of short hexa/octahistidine tags as opposed to fusion proteins.
- 4) Parallelization of small-scale expression screening in multiple hosts through the use of the pOPIN vector system.
- 5) Changing the host cell from *E.coli* when little or no expression is detected.
- 6) Use of standardised buffers and protocols during protein purification.

6. Acknowledgements

The OPPF was funded by the Medical Research Council, UK (grant MR/K018779/1). The authors wish to thank Louise Bird for help with data analysis.

7. References

1. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*. Bioinformatics, 2005. **21**(16): p. 3369-76.
2. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis*. Nat Protoc, 2015. **10**(6): p. 845-58.
3. Pajon, A., et al., *Design of a data model for developing laboratory information management and analysis systems for protein production*. Proteins, 2005. **58**(2): p. 278-84.
4. Morris, C., et al., *The Protein Information Management System (PiMS): a generic tool for any structural biology research laboratory*. Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 249-60.
5. Berrow, N.S., et al., *A versatile ligation-independent cloning method suitable for high-throughput expression screening applications*. Nucleic Acids Res, 2007. **35**(6): p. e45.
6. Berrow, N.S., D. Alderton, and R.J. Owens, *The precise engineering of expression vectors using high-throughput In-Fusion PCR cloning*. Methods Mol Biol, 2009. **498**: p. 75-90.
7. Bird, L.E., *High throughput construction and small scale expression screening of multi-tag vectors in Escherichia coli*. Methods, 2011. **55**(1): p. 29-37.
8. Bird, L.E., et al., *Application of In-Fusion cloning for the parallel construction of E. coli expression vectors*. Methods Mol Biol, 2014. **1116**: p. 209-34.
9. Zhao, Y., D.A. Chapman, and I.M. Jones, *Improving baculovirus recombination*. Nucleic Acids Res, 2003. **31**(2): p. E6-6.

10. Nettleship, J.E., et al., *Recent advances in the production of proteins in insect and mammalian cells for structural biology*. J Struct Biol, 2010. **172**(1): p. 55-65.
11. Aricescu, A.R., W. Lu, and E.Y. Jones, *A time- and cost-efficient system for high-level protein production in mammalian cells*. Acta Crystallogr D Biol Crystallogr, 2006. **62**(Pt 10): p. 1243-50.
12. Durocher, Y., S. Perret, and A. Kamen, *High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells*. Nucleic Acids Res, 2002. **30**(2): p. E9.
13. Nettleship, J.E., et al., *Transient expression in HEK 293 cells: an alternative to E. coli for the production of secreted and intracellular mammalian proteins*. Methods Mol Biol, 2015. **1258**: p. 209-22.
14. Brown, M.H. and A.N. Barclay, *Expression of immunoglobulin and scavenger receptor superfamily domains as chimeric proteins with domains 3 and 4 of CD4 for ligand analysis*. Protein Eng, 1994. **7**(4): p. 515-21.
15. Nettleship, J.E., N. Rahman-Huq, and R.J. Owens, *The production of glycoproteins by transient expression in Mammalian cells*. Methods Mol Biol, 2009. **498**: p. 245-63.
16. Chang, V.T., et al., *Glycoprotein structural genomics: solving the glycosylation problem*. Structure, 2007. **15**(3): p. 267-73.
17. Nettleship, J.E., et al., *Methods for protein characterization by mass spectrometry, thermal shift (ThermoFluor) assay, and multiangle or static light scattering*. Methods Mol Biol, 2008. **426**: p. 299-318.
18. Nettleship, J.E., et al., *Converting monoclonal antibodies into Fab fragments for transient expression in mammalian cells*. Methods Mol Biol, 2012. **801**: p. 137-59.
19. Joshi, H.J. and R. Gupta, *Eukaryotic Glycosylation: Online Methods for Site Prediction on Protein Sequences*, in *Glycoinformatics*, T. Lütke and M. Frank, Editors. 2015, Springer New York: New York, NY. p. 127-137.
20. Steentoft, C., et al., *Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology*. Embo j, 2013. **32**(10): p. 1478-88.

21. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
22. Gasteiger, E., et al., *Protein Identification and Analysis Tools on the ExPASy Server*, in *The Proteomics Protocols Handbook*, J.M. Walker, Editor 2005, Humana Press: Totowa, NJ. p. 571-607.
23. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions*. Nat Methods, 2011. **8**(10): p. 785-6.
24. Moller, S., M.D. Croning, and R. Apweiler, *Evaluation of methods for the prediction of membrane spanning regions*. Bioinformatics, 2001. **17**(7): p. 646-53.
25. *UniProt: the universal protein knowledgebase*. Nucleic Acids Res, 2017. **45**(D1): p. D158-d169.

Table Legends

Table 1: Web-based resources for bioinformatic analysis of protein sequences used routinely in the OPPF.

Table 2: Comparison of *E. coli* versus insect cell expression for constructs with N- and C-terminal His tags.

Table 1

NAME	USE	URL	REFERENCE
NetNGlyc	N-glycosylation predication	http://www.cbs.dtu.dk/services/NetNGlyc/	[19]
NetOGlyc	O-glycosylation prediction	http://www.cbs.dtu.dk/services/NetOGlyc/	[20]
PDB	Protein structure database	http://www.rcsb.org/	[21]
Phyre2	Structure homology model	http://www.sbg.bio.ic.ac.uk/phyre2/	[2]
ProtParam	Physical and chemical parameters	http://web.expasy.org/protparam/	[22]
RONN	Disorder predication	https://www.strubi.ox.ac.uk/RONN	[1]
SignalP	Signal sequence predication	http://www.cbs.dtu.dk/services/SignalP/	[23]
TMHMM	Transmembrane predication	http://www.cbs.dtu.dk/services/TMHMM/	[24]
Uniprot	Protein sequence and function information	http://www.uniprot.org/	[25]

Table 2

	pOPINE	pOPINF	pOPINE/F
Expression in either system	94	113	207
The same expression level	13	11	24
Better expression in <i>E. coli</i>	17	17	34
Better expression in insect cells	64	85	149

Figure Legends

Figure 1: Overview of the protein production pipeline.

Figure 2: Screenshot of output from RONN v3.2 showing disorder prediction. The solid line represents the probability of disorder for each amino acid with the dotted line showing the boundary between order and disorder. The amino acid sequence is given below the graph. In this case, an area of disorder can be seen at the C-terminus where the blue line is above the red dotted line for a string of amino acids. Two further short sequences of disorder can be seen within the protein sequence.

Figure 3: Diagrammatic representation of A) pOPINF and B) pOPINTTGneo. pOPINF is a vector conferring an N-terminal hexahistidine tag followed by a 3C protease cleavage site onto a gene of interest and is based on TriEx 1.1. pOPINTTGneo contains the RTPT μ signal sequence at the N-terminus of the gene of interest and a C-terminal histidine tag and is based on the pTT vector. (Maps generated using SnapGene Viewer)

Figure 4: pOPIN vector configurations. A) N-terminal tags. Vectors containing a hexahistidine tag (green) in combination with different fusion proteins (blue) and a 3C protease cleavage site (black) upstream of the protein-of-interest (red). B) C-terminal tags. The top two vectors contain the protein-of-interest (red) followed by a lysine before the hexahistidine tag (green). The lysine allows the tag to be removed by carboxypeptidase A. The second of these two vectors contains a signal sequence for periplasmic expression in *E. coli* (orange). The bottom four vectors contain a 3C protease cleavage site (black) after the protein-of-interest (red) and then longer C-terminal fusion proteins (blue) before the hexahistidine tag (green). C) Vectors based on pTT with the RTPT μ signal sequence (purple) for secreted expression in eukaryotic systems. pOPIN vector sequences are available on the OPPF website (www.oppf.rc-harwell.ac.uk).

Figure 5: Cloning and expression screening pipelines for intracellular proteins using *E. coli* and insect cells.

Figure 6: SDS-PAGE gel after NiNTA magnetic bead purification showing expression of constructs for DDR1 in A) *E. coli* and B) insect cells via the baculovirus system. The lanes are numbered

according to the construct amino acid start and stop points along with the vector which is either pOPINE “E”, pOPINF “F” or pOPINJ “J” conferring a C-terminal His tag, an N-terminal His tag or an N-terminal His-GST tag respectively. Triangles represent the expected molecular weight of the protein.

Figure 7: SDS-PAGE gel after NiNTA magnetic bead purification showing expression in A) *E. coli* and B) insect cells with 8 different N-terminal tags. The solubility tag is shown above the gel.

Expected molecular weights are as follows: His (pOPINF) = 38.5 kDa; SUMO – small ubiquitin-like modifier (pOPINS3C) = 49.5 kDa; NusA – N-utilisation substance protein A (pOPINNusA) = 93 kDa; MysB (pOPINMysB) = 52.5 kDa; MBP – maltose binding protein (pOPINM) = 79 kDa; TF – trigger factor (pOPINTF) = 86.5 kDa; TRX – thioredoxin (pOPINTRX) = 50 kDa; Halo7™ (pOPINHalo7) = 71.5 kDa. Triangles represent the expected molecular weight of the protein.

Figure 8: Western blot analysis of secreted expression in HEK 293 cells for human epididymis protein 4 (HE4) with the native signal peptide (pOPINEneo), the RTPT μ signal peptide (pOPINTTGneo) and the RTPT μ signal peptide and a C-terminal CD4 tag (pOPINTTGneo-CD4).

Figure 9: Pipeline for protein purification for A) intracellular proteins from *E. coli* and insect cells and B) secreted proteins from mammalian cells. 3C protease and EndoF1 are made in-house (3C protease clone from EMBL Heidelberg, Germany and EndoF1 clone from Weizmann Institute, Israel)

Figure 10: Size exclusion profiles from standardised purification of 12 products performed over a one week period. Proteins are labelled according to OPPF number and the host cell used for production is indicated. Arrows indicate the peak containing the protein of interest.

Figure 11: Intact protein mass spectrometry deconvoluted spectra from purified proteins. Proteins are labelled according to their OPPF number and host cell used for production is indicated.

Glycoproteins which have been treated with PNGase F to remove the glycans prior to intact protein mass spectrometry are labelled “PNGaseF”.

Figure 12: Results from a thermal shift assay investigating the effect of pH on protein stability for a human histone acetyltransferase. The assay is performed in 96 well format and only a sub-set of data is shown here.