Article

Biotechnology and Bioengineering

Performance Benchmarking of Four Cell-Free Protein Expression Systems

Dejan Gagoski, Mark E. Polinkovsky, Sergey Mureev, Anne Kunert, Wayne Johnston, Yann Gambin, Kirill Alexandrov

Institute for Molecular Bioscience, University of Queensland, St. Lucia 4072, QLD, Australia; telephone: +61-7-3346-2017; fax: +61-7-3346-2101; e-mail: k.alexandrov@uq.edu.au

ABSTRACT: Over the last half century, a range of cell-free protein expression systems based on pro- and eukaryotic organisms have been developed and have found a range of applications, from structural biology to directed protein evolution. While it is generally accepted that significant differences in performance among systems exist, there is a paucity of systematic experimental studies supporting this notion. Here, we took advantage of the species-independent translation initiation sequence to express and characterize 87 N-terminally GFP-tagged human cytosolic proteins of different sizes in *E. coli*, wheat germ (WGE), HeLa, and *Leishmania*-based (LTE) cell-free systems. Using a combination of single-molecule fluorescence spectroscopy, SDS-PAGE, and Western blot analysis, we assessed the expression yields, the fraction of full-length translation product, and aggregation propensity for each of these systems. Our results demonstrate that the *E. coli* system has the highest expression yields. However, we observe that high expression levels are accompanied by production of truncated species—particularly pronounced in the case of proteins larger than 70 kDa. Furthermore, proteins produced in the *E. coli* system display high aggregation propensity, with only 10% of tested proteins being produced in predominantly monodispersed form. The WGE system was the most productive among eukaryotic systems tested. Finally, HeLa and LTE show comparable protein yields that are considerably lower than the ones achieved in the *E. coli* and WGE systems. The protein products produced in the HeLa system display slightly higher integrity, whereas the LTE-produced proteins have the lowest aggregation propensity among the systems analyzed. The high quality of HeLa- and LTE-produced proteins enable their analysis without purification and make them suitable for analysis of multi-domain eukaryotic proteins.

Biotechnol. Bioeng. 2015;9999: 1–9.

© 2015 Wiley Periodicals, Inc.

KEYWORDS: in vitro protein translation; cell-free protein expression; protein aggregation; *E. coli* CF; WGE; HeLa; LTE; *Leishmania*

Introduction

Cell-free protein expression is a rapid and high-throughput methodology for conversion of genetic information into protein-mediated biochemical activities. It relies on the autonomous nature of cellular protein translation machinery that retains sufficient biosynthetic activity upon cell disruption and fractionation (Rosenblum and Cooperman, 2013). When engineered to include phage gene transcription machinery, cell-free systems can be primed with circular or linear DNA. Cell-free protein expression holds the promise of narrowing the technological gap between DNA and protein technologies and provides a platform for broad application of synthetic biology's principles in life sciences. Cell-free expression systems gained renewed attention recently with the explosive growth in uncharacterized open reading frames generated by genomic sequencing projects and the consequent need for rapid and scalable methods of protein production and analysis. The oldest and the best characterized cell-free system is based on *E. coli* (Nirenberg and Matthaei, 1961). Inexpensive preparation and high expression yields enable it to be used broadly, including such challenging applications as industrial-scale production of pharmaceutical proteins (Zawada et al., 2011). However, in keeping with the differences in pro- and eukaryotic biology, the *E. coli* cell-free system often generates eukaryotic multidomain proteins in truncated or aggregated form (Hillebrecht and Chong, 2008). Therefore, solubility tags have been used to reduce product aggregation (Aceti et al., 2015; Ahn et al., 2011) and the incubation temperature is lowered (Terada et al., 2014) to increase the proportion of full-length products. In many cases, template-specific optimization has been performed in order to yield a soluble and full-length product (Ishihara et al., 2005; Shimono et al., 2009). In effort to find a systematic solution to the challenges of the *E. coli* cell-free system, as well as to create systems capable of eukaryotic-type posttranslational modifications, a range of alternative cell-free systems based on protozoa, yeast, plants, insects, and mammals have been developed (Rosenblum and Cooperman, 2013).

Contract grant sponsor: ARC DP
Contract grant number: DP1094080
Contract grant sponsor: ARC FF
Contract grant numbers: FT0991611; LP120100162
Contract grant sponsor: NHMRC
Contract grant number: 569652
Contract grant sponsor: NHMRC
Contract grant number: APP1037320
Received 16 April 2015; Revision received 20 July 2015; Accepted 19 August 2015
Accepted manuscript online xx Month 2015;
Article first published online in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/bit.25814

The eukaryotic systems that became widespread and commercially available include wheat germ extract (WGE), rabbit reticulocyte lysate (RRL), insect cell extract (ICE), HeLa cell extract, and the *Leishmania tarentolae* cell-free system (LTE), with several more systems reported (Buntru et al., 2014; Wang et al., 2014; Zhou et al., 2012). Although many systems co-exist and have been used successfully in different applications, limited data comparing the systems' performance are available. A systematic analysis of artificial protein domain fusions demonstrated that an unoptimized *E. coli* system is less efficient in folding multidomain proteins than the WGE (Hirano et al., 2006). Another study comparing *E. coli*, WGE, and HeLa extract came to similar conclusions (Saul et al., 2014). These studies also used relatively small experimental sets that may skew the results in favor of one or another system. While it is generally accepted that pro- and eukaryotic cell-free systems differ in their ability to produce multidomain proteins in functional form, the possible differences among eukaryotic systems are much less clear. It is unknown whether some systems are better suited for expression of a particular type of protein or whether the performance differs only along the pro-/eukaryotic divide. Performance comparison of cell-free systems is complicated by the fact that sequences mediating translational initiation are traditionally optimized for a particular system and perform poorly in other systems. Furthermore, large differences exist between translation initiation mechanisms in pro- and eukaryotes, further complicating the engineering of translation initiation sequences with broad specificity (Kozak, 1999). We have recently addressed this problem by creating a polymeric sequence that can promote the assembly of translational complexes, presumably by bypassing early translation initiation factors (Mureev et al., 2009). The sequence was termed species-independent translation initiation sequence (SITS) and was demonstrated to initiate efficient translation in eukaryotic cell-free systems (Mureev et al., 2009), as well as in the prokaryotic *E. coli* cell-free system, when combined with an upstream $(UUUA)_{11}$ sequence (Kovtun et al., 2010). Subsequently, we used the SITS to construct a suite of Gateway-compatible vectors that allow rapid cloning and in vitro expression of open reading frames (Gagoski et al., 2015). In the presented study, we take advantage of the universal nature of the SITS-based vectors to compare four cell-free systems using a representative set of proteins. The set was produced in batch using the transcription/translation coupled mode as the most streamlined and high-throughput configuration. To get an insight into the global expression characteristics of the four systems, we avoided the use of cell-free system versions optimized for specific sets of proteins. In order to benchmark the cell-free systems, we chose to analyze three main parameters of the recombinant polypeptides: expression levels, product integrity, and the aggregation propensity. We observe striking differences in the abilities of the cell-free systems to produce full-length monodisperse proteins.

Materials and Methods

Construction of the Expression Library

The initial list of candidate ORFs was produced by ranking human proteins according to the number of hits in the Google search engine. The resulting list was then binned according to the predicted molecular weights of the proteins and matched with the clones in Human ORFeome collection version 5.1 (Open Biosystems) or the ImaGenes ORFeome set (Source Bio-science, Lifesciences), compiled as described earlier (Škalamera et al., 2011). A total of 68 out of the 96 candidates were available in this clone collection and were successfully cloned in the pCellFree G03 vector in frame with the N-terminal GFP as described previously (Gagoski et al., 2015). Furthermore, 18 candidate proteins, which were of high research interest to our group, were also cloned in the same vector (Table S1, Supplementary Data), and the resulting library and its parental vectors were deposited to the public AddGene collection (Cell-free expression test kit). The resulting 87 expression constructs were amplified by plasmid mini preparation with the QIAprep Spin Miniprep Kit (Qiagen) or midi preparation with the NucleoBond® Xtra Midi/Maxi kit (Macherey-Nagel) and diluted to a final concentration of 130 nM. In order to control for the differences of plasmid quality of different DNA preparation methods, the STAT3 construct was represented twice, once as a mini- and once as a midi-prepared plasmid.

Cell-Free Expression Systems and In Vitro Protein Expression

The *E. coli*-based cell-free expression system (*E. coli* CF) was prepared as described previously (Schwarz et al., 2007). For this, a colony of BL21 (DE23) Gold from an LB-agar plate with tetracycline was used to inoculate 50 mL of LB. The starting culture was grown overnight at 37°C and used to inoculate 5 L of TBGG medium (12 g/L Trypton, 24 g/L Yeast extract, 0.8% Glycerol, 5.55 mM Glucose, 17 mM KH_2PO_4, 72 mM K_2HPO_4). This culture was grown to an $OD_{600} = 4.5$ at 37°C and cooled down using 1 L of LB frozen at −80°C. Cells were then harvested by centrifugation for 15 min at 2500 g and washed twice with 1 L of S30A buffer. The cell pellet was resuspended in 200% v/w (mL/g) S30B buffer (10 mM Tris-acetate pH 8.2, 14 mM $Mg2^+$ acetate, 0.6 mM KCl, 1 mM DTT, 0.1 mM PMSF). The cells were then disrupted using a fluidiser (Constant Systems) at 20 kpsi, and the S30 extract was separated from the cell debris by a 30-min centrifugation at 30,000 g. The supernatant was adjusted with 0.4 M NaCl and incubated at 42°C for 45 min. After the incubation, the lysate was dialysed against 5 L of cold S30C buffer (10 mM Tris-acetate (pH 8.2), 14 mM Mg2+ acetate, 0.6 mM K+ acetate, 0.5 mM DTT) for 2 h and overnight against another 5 L of the same buffer. The lysate was finally centrifuged again at 30,000 g for 30 min, aliquoted, and snap-frozen in liquid nitrogen. For expression of the library constructs, lysate was mixed with feed solution (2% PEG 8000; 148 mM K-acetate; 5 mM Mg-acetate; 86 mM HEPES-KOH pH 8.0; 1× Roche protease inhibitor mix; 0.1 mg/mL folic acid; 2 mM DTT; 1.2 mM ATP; 0.8 mM GTP, CTP, UTP each; 15 mM acetate-PO_4; 24 mM creatine phosphate; 0.5 mM of each 20 amino acids; 1 mM of each R, C, W, D, M, and E amino acids; 0.045 U/μL creatine phosphokinase; 0.1 mg/mL T7 polymerase) and plasmid DNA, mixed in a 10 lysate:13 feed solution:5 DNA v/v ratio and incubated at 30°C. The wheat germ-based cell-free system was prepared by mixing the WEPRO1240 lysate (CellFree Sciences) with the feed solution (1.7 mM ATP; 0.634 mM GTP; 0.25 mM spermidine; 2 mM DTT; 40 mM creatine phosphate; 8 mM HEPES KOH pH 7.6; 0.01 mM PEG3000; 1×

Roche protein inhibitor mix; 0.136 of each 20 amino acids; 0.5 mM UTP; 0.5 mM CTP; 0.04 U/μL creatine phosphokinase; 0.1 mg/mL T7 polymerase; 0.01 mM oligonucleotide), 80 mM Mg-acetate and plasmid DNA in a 10 lysate:5 feed solution:1 Mg-acetate:4 DNA v/v ratio as described previously (Kovtun et al., 2011). The reactions were incubated at 30°C. For expression in the HeLa cell-based cell-free system, the 1-Step Human Coupled IVT kit (Thermo Fisher Scientific) was used. The reactions were prepared by mixing HeLa cell lysate, accessory proteins, reaction mix and plasmid DNA in 10:2:4:4 v/v ratios and incubated at 30°C. The lysate and feed solution for the LTE system were prepared as previously described (Johnston and Alexandrov, 2014), and the reactions were set up by mixing *L. tarentolae* lysate, feed solution (containing 12 mM Mg-acetate) and DNA in a 10:6:4 ratio. The LTE expression reactions were incubated at 27°C. All systems were supplemented with RNaseOUT (Life Technologies) at 1:100 v/v to inhibit possible RNase contamination from the plasmid DNA preparations. The expression of the library was performed using 384-well plates in a Synergy 4 plate reader (BioTek), where the progress of the translation reactions was monitored through GFP fluorescence. The total protein concentrations of the lysates were measured using the Pierce™ BCA Protein Assay Kit (Thermo Fischer Scientific) and the Bradford-based Bio-Rad Protein Assay (Bio-Rad).

Expression Levels and Aggregation Propensity

The eGFP fluorescence levels of the samples were measured using a custom-built single-molecule analysis system based on a Zeiss Axio Observer microscope. Each sample was excited by a 488-nm solid state laser (iBeam Smart, Toptica Photonics) through a Zeiss 40×/1.2 WI objective, and the fluorescence signals of the eGFP tags were detected in a confocal volume, defined by the 50-μm-wide pinhole of an avalanche photodiode detector (τ-SPAD, PicoQuant GmbH). A portion of each sample was diluted in 10 mM, pH 7.2 HEPES buffer by a factor calculated for each cell-free system to ensure that the GFP fluorescence was at a comparable level and that the fluorescence signal would not saturate the detector. For each construct, the fluorescence signal was recorded for 30 s with 1 ms binning, resulting in 30,000 measurements per sample. Subsequently, the mean, standard deviation and skewness of each sample's signal distribution were calculated. The mean signal was then correlated with a calibration curve obtained using purified eGFP. As an independent measure of protein concentration, the samples were analyzed by Western blotting (Supplementary Information, Figure S9).

Integrity Analysis of the In Vitro-Expressed Proteins

The integrity of the expressed products was analyzed and quantified by fluorescent scanning of SDS-PAGE loaded with the corresponding translation reaction in order to establish the ratio between the full-length protein and small products. For that, 2.5–5-μL aliquots of the expression reactions were mixed 1:1 with NuPAGE® LDS Sample Buffer (Life Technologies) and incubated at 72°C for 150 s. A total of 4 μL of the resulting sample was loaded on 15-well NuPAGE® Novex® 4–12% Bis–Tris Protein Gels, 1.0 mm (Life Technologies), alongside 5 μL of PageRuler plus pre-stained protein ladder (Thermo Fisher Scientific). The fluorescence of the GFP-tagged products was detected with the ChemiDoc™ MP System (Bio Rad), and the lane and band analysis was performed with the Image Lab™ software (Bio Rad).

Results and Discussion

Expression Library Design and Construction

In order to analyze the performance differences of cell-free systems, we chose to analyze extracts prepared from the most distantly related organisms: *E. coli,* WGE, HeLa, and the recently developed LTE system. These systems were benchmarked on the test protein set of 87 ORFs from the human Gateway ORFeom library (Lamesch et al., 2007; Rual et al., 2004), selected based on their calculated molecular weight and the citation frequency. To cover different size ranges, we binned ORFs in to 160–90 kDa, 90–60 kDa, 60–50 kDa, 50–40 kDa, 40–25 kDa, and 25–4 kDa groups with 13–17 proteins per bin (Table S1). Out of the 87 ORFs, 19 have not been previously expressed as recombinant proteins, 51 have been expressed in *E. coli*, 14 have been expressed in HeLa cells, 22 have been expressed in other eukaryotic expression systems, and 10 have been expressed in cell-free systems (Table S2). These ORFs were cloned into the pCellFree-G03 vector, featuring a SITS sequence followed by the eGFP gene and the recombination sites for Gateway-based ORF cloning. The transfer of the chosen ORFs into this vector was achieved by parallelized recombination Gateway cloning as

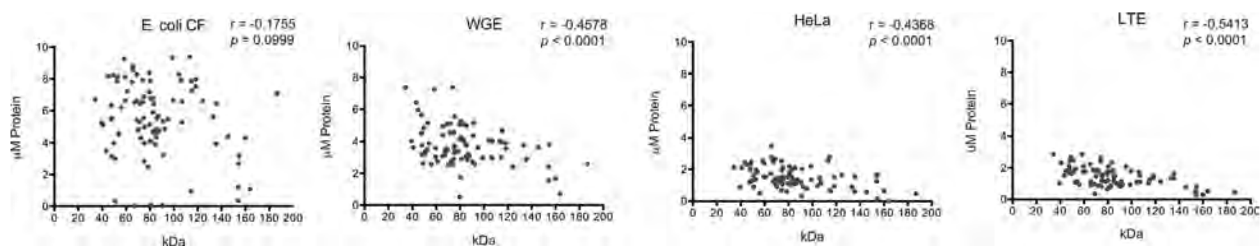


Figure 1. Expression levels of the individual constructs from the test library plotted against the protein size. The *E. coli* and the WGE systems, on average, display significantly higher levels of expression in comparison to the HeLa, IVT, and LTE systems. Eukaryotic cell-free systems display a strong correlation between the expression level and the length of the protein. The expression levels of the individual constructs were calculated both in μM and in μg/mL (Figure S2, Table S3).

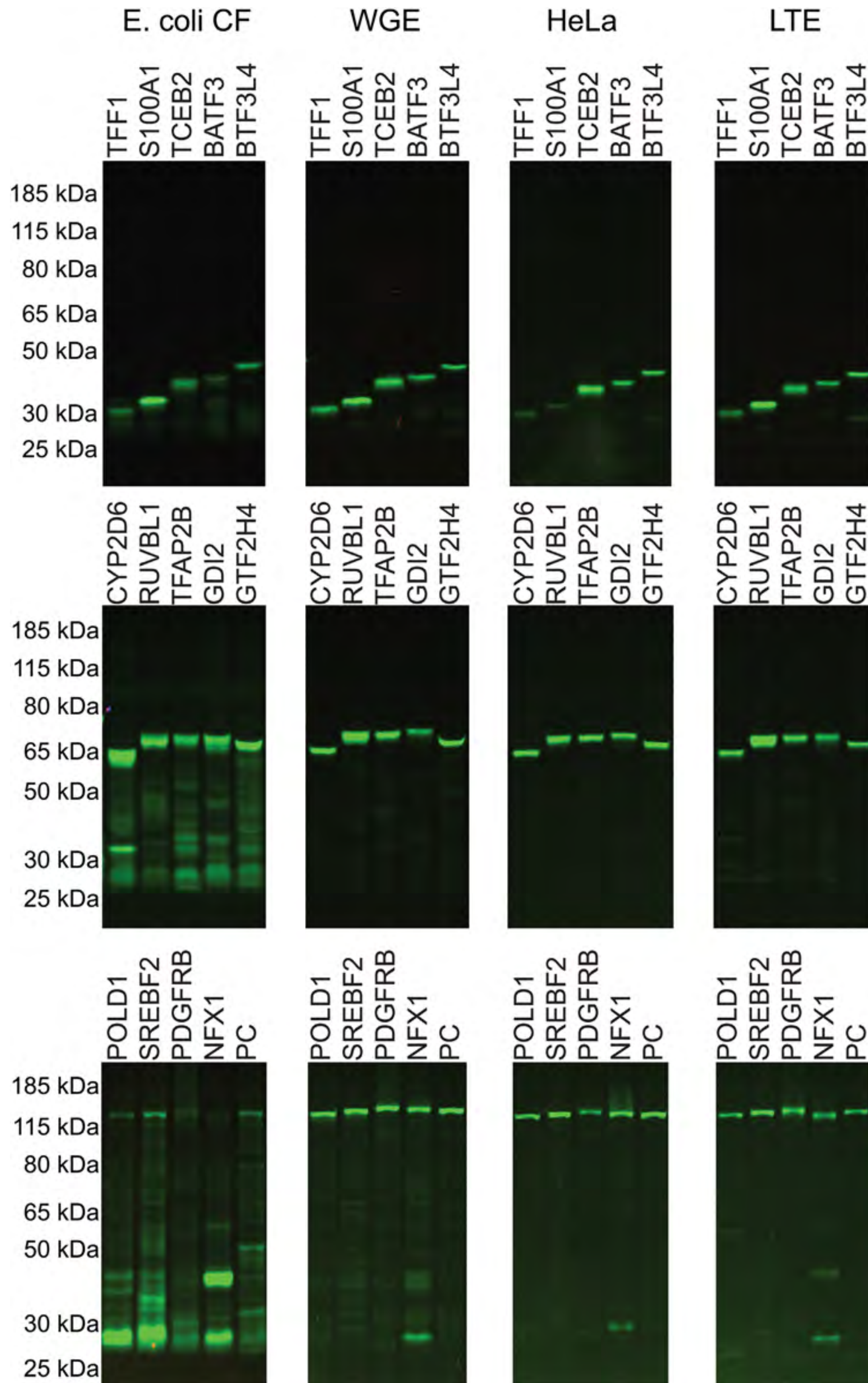


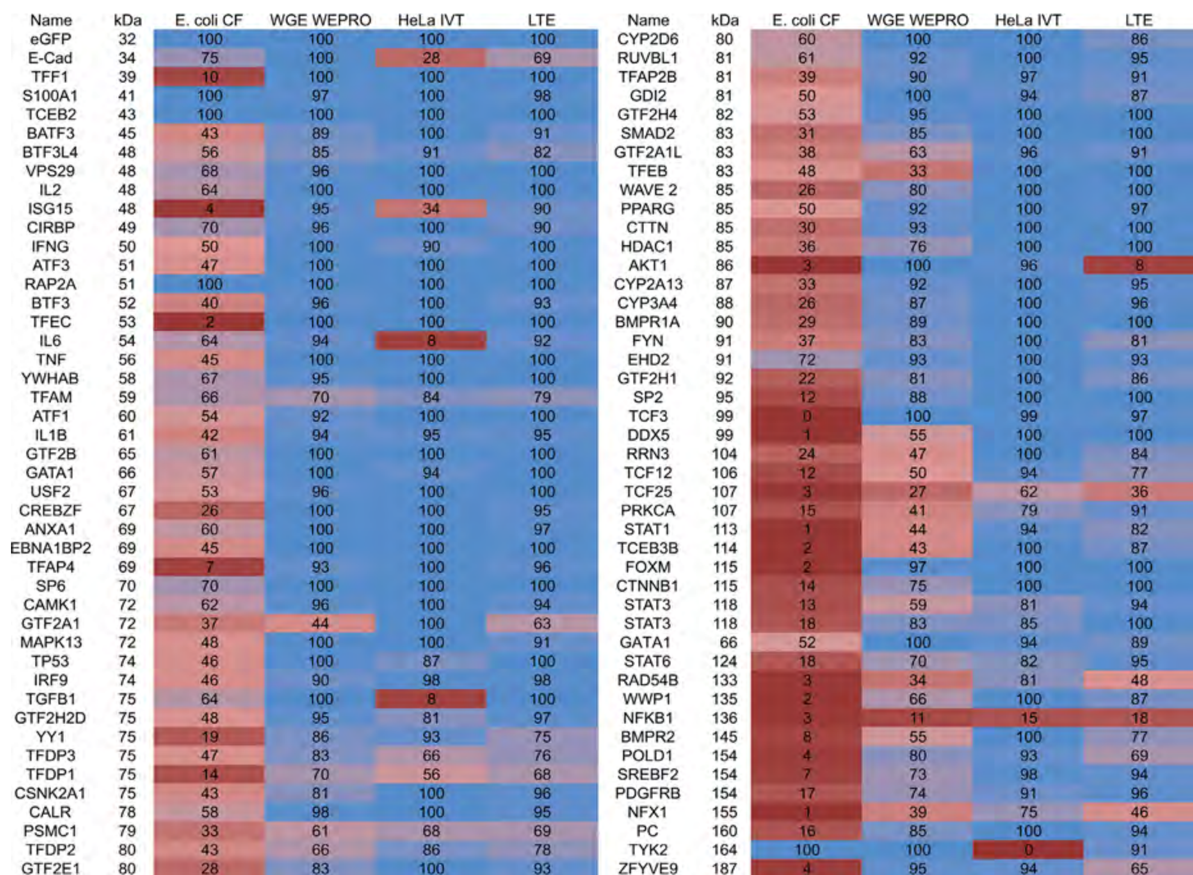
Figure 2. Examples of SDS-PAGE analysis of in vitro-translated N-terminally GFP-tagged proteins of different sizes in different translation systems. The fluorescence of the full-length and the truncated products was quantified and used to assess the integrity of the proteins.

described previously (Škalamera et al., 2011). The DNA of the resulting clones was isolated, and the identity of the clones was confirmed by sequencing.

Comparing the Expression Levels Among Cell-Free Systems

We expressed the test library in all four cell-free systems and followed the expression kinetics using GFP fluorescence as a reporter. The reactions were analyzed immediately once the fluorescence signal was saturated. The level of protein expression was analyzed by recording the fluorescence of the eGFP tag in the diluted translation reactions using a confocal microscope (Fig. 1). We observed that on average, the *E. coli* CF produces the highest amount of fluorescence, followed by the WGE system, with 68% of the *E. coli* yields. Both HeLa and the LTE show significantly lower expression levels, producing 27.8% and 25.4% of *E. coli*'s yield, respectively. These results are in agreement with previous observations (Chong, 2014; Endo and Sawasaki, 2006) that the *E. coli* cell-free system is significantly more productive than the eukaryotic cell-free systems. In our analysis, the expression yield was normalized to reaction volume, which provides no information on the concentration of the translational machinery. As another possible measure of productivity, we use the relationship between the total protein concentration and the expression yield of the system (Figure S3). We observed that the prokaryotic system requires lower lysate-protein concentrations than the eukaryotic system for higher product yields. This increased productivity of the prokaryotic system is likely to be a consequence of simpler translational machinery.

Our results indicate that constructs of similar sizes may be expressed at different levels in different systems. This may be caused by codon bias of the templates and the secondary structures formed by their mRNA, as well as other less-characterized factors. Nevertheless, the availability of a large dataset enabled us to establish the correlation between expression levels and the product size. We detected a strong negative correlation between the size and the expression yield for all three eukaryotic cell-free systems (P-values ≤ 0.0001), whereas the prokaryotic system did not show a significant correlation (P-value $= 0.1615$). This may be explained by the fact that more substrate (amino acids, rNTPs) and energy (ATP and GTP) are needed to synthesize larger molecules. This suggests that depletion of the substrates may be limiting the productivity of the tested eukaryotic systems. On the other hand, there is a positive correlation between the expressed protein mass and the product size (Figure S2). This suggests that irrespective of the system, the substrates and the translation machinery are utilized more efficiently for longer products. This discrepancy between the expressed molar and mass concentrations present in



Name	kDa	E. coli CF	WGE WEPRO	HeLa IVT	LTE
eGFP	32	100	100	100	100
E-Cad	34	75	100	28	69
TFF1	39	10	100	100	100
S100A1	41	100	97	100	98
TCEB2	43	100	100	100	100
BATF3	45	43	89	100	91
BTF3L4	48	56	85	91	82
VPS29	48	68	96	100	100
IL2	48	64	100	100	100
ISG15	48	4	95	34	90
CIRBP	49	70	96	100	90
IFNG	50	50	100	90	100
ATF3	51	47	100	100	100
RAP2A	51	100	100	100	100
BTF3	52	40	96	100	93
TFEC	53	2	100	100	100
IL6	54	64	94	8	92
TNF	56	45	100	100	100
YWHAB	58	67	95	100	100
TFAM	59	66	70	84	79
ATF1	60	54	92	100	100
IL1B	61	42	94	95	95
GTF2B	65	61	100	100	100
GATA1	66	57	100	94	100
USF2	67	53	96	100	100
CREBZF	67	26	100	100	95
ANXA1	69	60	100	100	97
EBNA1BP2	69	45	100	100	100
TFAP4	69	7	93	100	96
SP6	70	70	100	100	100
CAMK1	72	62	96	100	94
GTF2A1	72	37	44	100	63
MAPK13	72	48	100	100	91
TP53	74	46	100	87	100
IRF9	74	46	90	98	98
TGFB1	75	64	100	8	100
GTF2H2D	75	48	95	81	97
YY1	75	19	86	93	75
TFDP3	75	47	83	66	76
TFDP1	75	14	70	56	68
CSNK2A1	75	43	81	100	96
CALR	78	58	98	100	95
PSMC1	79	33	61	68	69
TFDP2	80	43	66	86	78
GTF2E1	80	28	83	100	93

Name	kDa	E. coli CF	WGE WEPRO	HeLa IVT	LTE
CYP2D6	80	60	100	100	86
RUVBL1	81	61	92	100	95
TFAP2B	81	39	90	97	91
GDI2	81	50	100	94	87
GTF2H4	82	53	95	100	100
SMAD2	83	31	85	100	100
GTF2A1L	83	38	63	96	91
TFEB	83	48	33	100	100
WAVE 2	85	26	80	100	100
PPARG	85	50	92	100	97
CTTN	85	30	93	100	100
HDAC1	85	36	76	100	100
AKT1	86	3	100	96	8
CYP2A13	87	33	92	100	95
CYP3A4	88	26	87	100	96
BMPR1A	90	29	89	100	100
FYN	91	37	83	100	81
EHD2	91	72	93	100	93
GTF2H1	92	22	81	100	86
SP2	95	12	88	100	100
TCF3	99	0	100	99	97
DDX5	99	1	55	100	100
RRN3	104	24	47	100	84
TCF12	106	12	50	94	77
TCF25	107	3	27	62	36
PRKCA	107	15	41	79	91
STAT1	113	1	44	94	82
TCEB3B	114	2	43	100	87
FOXM	115	2	97	100	100
CTNNB1	115	14	75	100	100
STAT3	118	13	59	81	94
STAT3	118	18	83	85	100
GATA1	66	52	100	94	89
STAT6	124	18	70	82	95
RAD54B	133	3	34	81	48
WWP1	135	2	66	100	87
NFKB1	136	3	11	15	18
BMPR2	145	8	55	100	77
POLD1	154	4	80	93	69
SREBF2	154	7	73	98	94
PDGFRB	154	17	74	91	96
NFX1	155	1	39	75	46
PC	160	16	85	100	94
TYK2	164	100	100	0	91
ZFYVE9	187	4	95	94	65

Figure 3. Integrity of recombinant proteins produced in four cell-free systems represented as a heat map. High values (blue) reflect the high proportion of the full-length protein, while low values (red) indicate the reverse. The data are scaled to 100%.

the eukaryotic cell-free systems implies the interplay between the mRNA transcription rates and stability on one hand and the proportion of tRNA-unbound and -bound ribosomes, initiation, and elongation factors on the other hand. In addition, the fact that in the *E. coli* system, the product concentration does not correlate with its size may indicate that here, protein synthesis terminates before the primary substrates are depleted. This idea finds indirect support in the observation that the *E. coli* system reaches 90% of its maximum production yield almost two times faster than the eukaryotic systems (Figure S1). This may also reflect depletion of the template due to reduction in the efficiency of the transcription–translation reaction (Shin and Noireaux, 2010) or by an overall change in the reaction environment, like pH or macromolecular crowding (Tan et al., 2013).

Analysis of Protein Integrity

In vivo protein integrity is controlled by multiple mechanisms mainly linked to chaperone and degradation machineries (Benyair et al., 2011; Tasaki et al., 2012). In cell-free expression systems, these mechanisms are expected to be only partially operational. Furthermore, the disruption of the intracellular compartments and the activation of stress response mechanisms may lead to an increase of proteolytic activities in the lysate (Barchinger and Ades, 2013; Fulda et al., 2010; Martinou et al., 2000). This makes the cell-free systems susceptible to translation of 3′-degraded mRNAs, release of incomplete translation products, and accumulation of cleaved or partially degraded proteins. Hence, as a second parameter in our benchmarking exercise, we assessed the integrity of the protein products. We made use of the fact that N-terminal GFP remains fluorescent on the SDS-PAGE following mild heat denaturation treatment of the sample (Fig. 2). We quantified the amount of full-length and truncated products by fluorescent scanning of the SDS-PAGE. The ratio of the signal of the full-length product relative to the total signal in the lane was taken as a measure of the product's integrity (Fig. 3).

After visual inspection of the gels (Figure S4–S7), it becomes apparent that the fraction of truncated products increases in all systems with the increase in protein size. For the *E. coli* system, this becomes pronounced already with proteins larger than 70 kDa, whereas in the WGE system, truncated products become prominent with protein sizes above 100 kDa. The HeLa system produces the lowest amount of truncated proteins, followed by the LTE system, which produces full-length proteins in more than 90% of the cases.

The semi-quantitative analysis of the gels clearly points to the difference in product integrity between the prokaryotic and the eukaryotic systems, where the eukaryotic systems more consistently produce full-length proteins (Fig. 3 and S8). This difference is possibly in part due to the *E. coli* ribosome pausing on eukaryotic genes (Ramachandiran et al., 2000). The integrity of the products >70 kDa in the *E. coli* system can possibly be improved by optimizing the lysate preparation protocol and the expression conditions (Matsuda et al., 2006; Terada et al., 2014). The high expression yields and scalability of the system may make such optimization worthwhile.

Analysing Aggregation Propensity of the Recombinant Fusion Proteins Produced in Different Cell-Free Systems

Protein aggregation is one of the main problems of recombinant protein expression and often reflects incomplete or non-native protein folding. Monodispersity in solution is one of the most

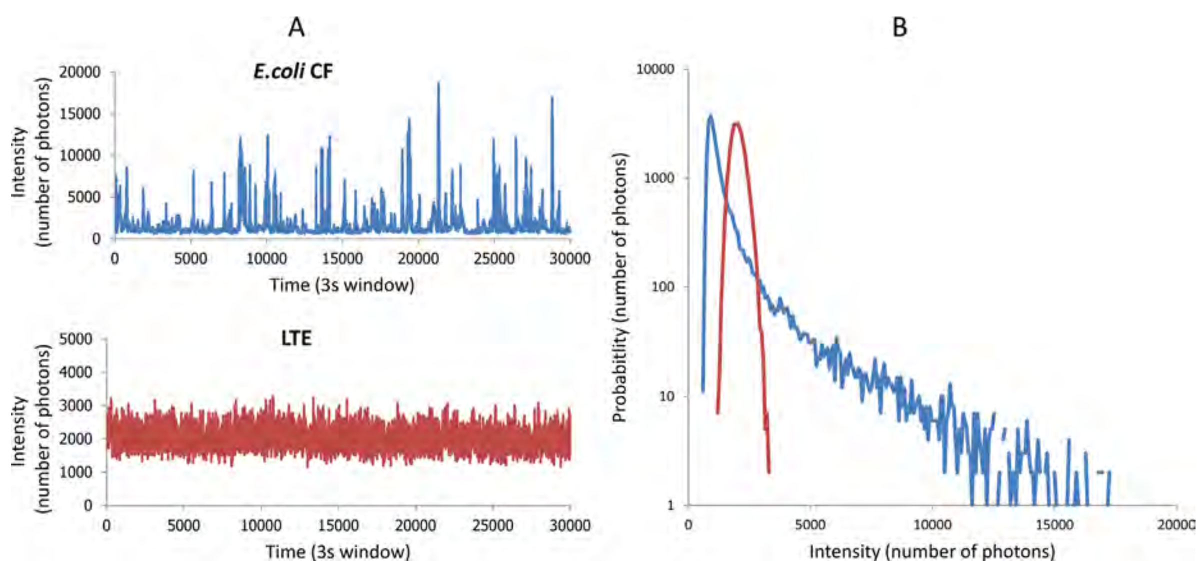
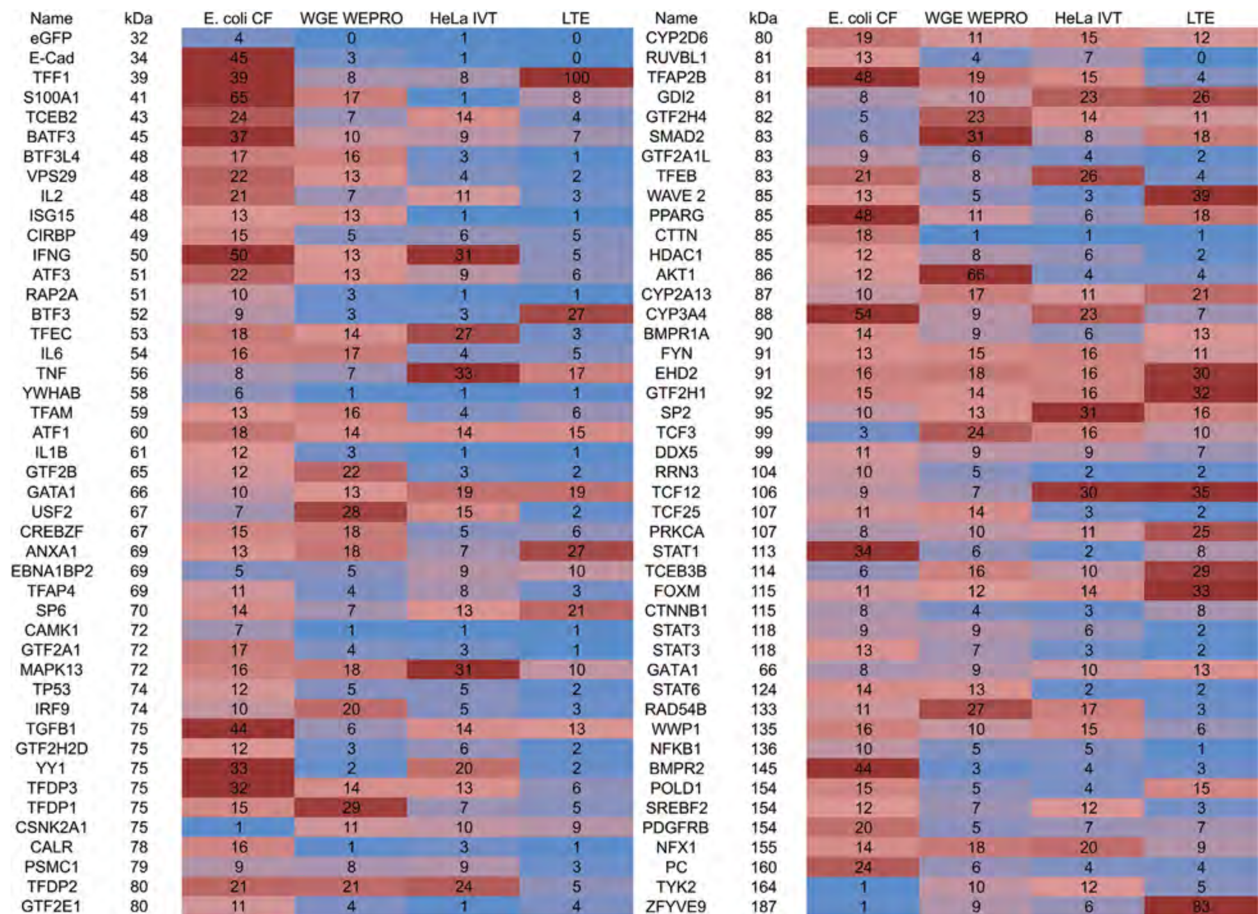


Figure 4. Single-molecule fluorescence analysis of protein aggregation of GFP-tagged VPS29 expressed in two cell-free systems. (**A**) Time trace of single-molecule fluorescence of the GFP-tagged VPS29 expressed in the *E. coli* CF system (blue trace) shows large spikes produced by aggregates containing large numbers of GFP-tagged proteins, while the trace for the same protein expressed in the LTE system (red trace) shows a monodispersal distribution of GFP-tagged proteins. (**B**) Frequency distribution of measured fluorescent intensities is symmetrical for LTE-expressed monodispersed VPS29 (red) and skewed for the *E. coli* CF-expressed aggregated VPS29 (blue) expression products. An increase in skewness reflects increased aggregation propensity.

commonly used parameters for assessing the quality of the recombinant proteins. We used single-molecule fluorescence analysis of GFP to assess the oligomerization state of the expressed polypeptides. We have previously demonstrated that single-molecule fluorescence of GFP can be used to estimate the size and dispersion of protein aggregates in solution (Gambin et al., 2014). We used this approach to record single-molecule fluorescence of each member of our expression library. We then analyzed the intensity of fluorescence bursts of each translation reaction by plotting frequency versus brightness for each sample. In the case of monodispersed proteins, this resulted in a Gaussian distribution, whereas in the presence of aggregates, the distribution is skewed toward higher brightness, as shown in the example of VPS29 protein (Fig. 4). A skewness index ≥ 1 is characteristic of samples with significant aggregation. This approach allowed us to rapidly assess the aggregation propensity of the test library (Fig. 5).

Inspection of the resulting heat map revealed the individual and systemic differences in protein aggregation. Unlike in the case of the protein integrity analysis, we did not observe a significant correlation between the product size and the aggregation propensity. This suggests that the sequence and the properties of the protein, as well as the properties of the expression system, determine the protein's aggregation propensity in cell-free expression. The distribution of the skewness values (Fig. 6) demonstrates that the products are much more likely to aggregate (values 1–4) in the prokaryotic than in eukaryotic cell-free systems. Interestingly, only ~10% of test proteins expressed in the *E. coli* system were fully monodispersed.

Out of the three eukaryotic systems, the LTE system showed the lowest propensity in forming protein aggregates. The significantly higher aggregation propensity of the *E. coli* system most probably reflects the high yields and the fast expression rate. It is also apparent from this analysis that the *E. coli* CF system produces, on average, larger protein aggregates than the eukaryotic cell-free systems (Fig. 6). The relatively high translation rate can lead to spatial proximity of the elongating nascent polypeptides, promoting their interaction and non-physiological aggregation. Thus, reducing the expression rate of the *E. coli* cell-free system or using a system based on a rare tRNA- and chaperone-expressing *E. coli* strain (Terada et al., 2014) could possibly reduce the higher aggregation propensity of the expressed proteins. We also considered the possibility that the truncated products present in some of the samples influenced our



Name	kDa	E. coli CF	WGE WEPRO	HeLa IVT	LTE
eGFP	32	4	0	1	0
E-Cad	34	45	3	1	0
TFF1	39	39	8	8	100
S100A1	41	65	17	1	8
TCEB2	43	24	7	14	4
BATF3	45	37	10	9	7
BTF3L4	48	17	16	3	1
VPS29	48	22	13	4	2
IL2	48	21	7	11	3
ISG15	48	13	13	1	1
CIRBP	49	15	5	6	5
IFNG	50	50	13	31	5
ATF3	51	22	13	9	6
RAP2A	51	10	3	1	1
BTF3	52	9	3	3	27
TFEC	53	18	14	27	3
IL6	54	16	17	4	5
TNF	56	8	7	33	17
YWHAB	58	6	1	1	1
TFAM	59	13	16	4	6
ATF1	60	18	14	14	15
IL1B	61	12	3	1	1
GTF2B	65	12	22	3	2
GATA1	66	10	13	19	19
USF2	67	7	28	15	2
CREBZF	67	15	18	5	6
ANXA1	69	13	18	7	27
EBNA1BP2	69	5	5	9	10
TFAP4	69	11	4	8	3
SP6	70	14	7	13	21
CAMK1	72	7	1	1	1
GTF2A1	72	17	4	3	1
MAPK13	72	16	18	31	10
TP53	74	12	5	5	2
IRF9	74	10	20	5	3
TGFB1	75	44	6	14	13
GTF2H2D	75	12	3	6	2
YY1	75	33	2	20	2
TFDP3	75	32	14	13	6
TFDP1	75	15	29	7	5
CSNK2A1	75	1	11	10	9
CALR	78	16	1	3	1
PSMC1	79	9	8	9	3
TFDP2	80	21	21	24	5
GTF2E1	80	11	4	1	4

Name	kDa	E. coli CF	WGE WEPRO	HeLa IVT	LTE
CYP2D6	80	19	11	15	12
RUVBL1	81	13	4	7	0
TFAP2B	81	48	19	15	4
GDI2	81	8	10	23	26
GTF2H4	82	5	23	14	11
SMAD2	83	6	31	8	18
GTF2A1L	83	9	6	4	2
TFEB	83	21	8	26	4
WAVE 2	85	13	5	3	39
PPARG	85	48	11	6	18
CTTN	85	18	1	1	1
HDAC1	85	12	8	6	2
AKT1	86	12	66	4	4
CYP2A13	87	10	17	11	21
CYP3A4	88	54	9	23	7
BMPR1A	90	14	9	6	13
FYN	91	13	15	16	11
EHD2	91	16	18	16	30
GTF2H1	92	15	14	16	32
SP2	95	10	13	31	16
TCF3	99	3	24	16	10
DDX5	99	11	9	9	7
RRN3	104	10	5	2	2
TCF12	106	9	7	30	35
TCF25	107	11	14	3	2
PRKCA	107	8	10	11	25
STAT1	113	34	6	2	8
TCEB3B	114	6	16	10	29
FOXM	115	11	12	14	33
CTNNB1	115	8	4	3	8
STAT3	118	9	9	6	2
STAT3	118	13	7	3	2
GATA1	66	8	9	10	13
STAT6	124	14	13	2	2
RAD54B	133	11	27	17	3
WWP1	135	16	10	15	6
NFKB1	136	10	5	5	1
BMPR2	145	44	3	4	3
POLD1	154	15	5	4	15
SREBF2	154	12	7	12	3
PDGFRB	154	20	5	7	7
NFX1	155	14	18	20	9
PC	160	24	6	4	4
TYK2	164	1	10	12	5
ZFYVE9	187	1	9	6	83

Figure 5. Aggregation propensity of the test protein library produced in different cell-free systems assessed by the single-molecule brightness analysis. The aggregation propensity was expressed as the skewness of the brightness distribution and scaled 0–100%, where 100% is the highest recorded skewness in the dataset. Low skewness levels (blue) represent products that do not form a considerable number of aggregates, whereas high skewness levels (red) are an indication of a large number of aggregates. The molecular weight of the GFP tag was included into the calculation of the molecular weights of the constructs.

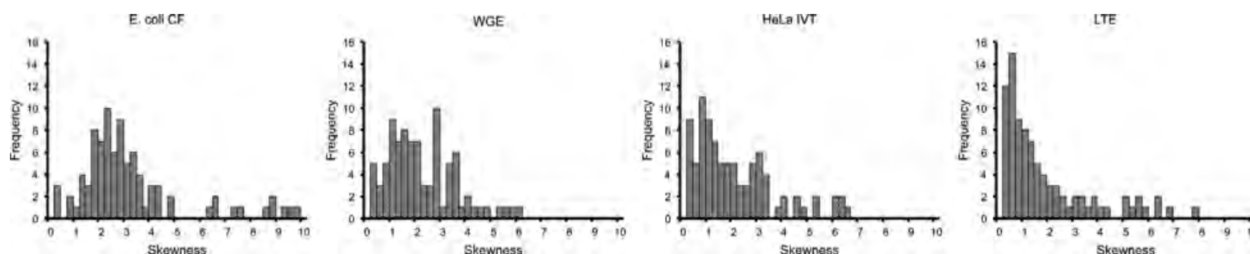


Figure 6. Aggregation propensity analysis of cell-free systems based on statistics of fluorescence signals of GFP-fusion proteins. The distributions of the skewness values produced in each cell-free system were plotted to compare their tendency to produce aggregating proteins. The *E. coli* CF makes the highest proportion of aggregating proteins, while the LTE system is least prone to protein aggregation.

analysis. However, no obvious correlation between protein integrity index and the aggregation propensity was observed.

Conclusions

Here, we present a benchmarking study of four leading cell-free systems on a size-selected set of human genes. The benchmarking included expression yield, product integrity, and aggregation propensity for each of 87 genes in the library. In keeping with previous studies, the analysis demonstrates that the *E. coli* system has the highest expression yields. However, we observe that high expression levels are accompanied by production of truncated species. This is particularly pronounced in the case of proteins larger than 70 kDa. Furthermore, proteins produced in the *E. coli* system display significantly higher aggregation propensity. This echoes the in vivo situation, where inclusion body formation in *E. coli* is a very common phenomenon.

The presented data were obtained with the most commonly used *E. coli* cell-free system, and its further optimization may improve its performance. The developed benchmarking set of expression constructs and their analysis methods described here provide the means for obtaining definitive answers to these questions. According to our analysis, the WGE system is the most productive among the eukaryotic systems tested. The product integrity is better than in the *E. coli* system, although the system does not handle proteins >100 kDa as well as the other tested eukaryotic systems. Finally, HeLa and LTE show comparable protein yields, although they are considerably lower than the ones achieved in the *E. coli* and WGE systems. The protein products produced in the HeLa system display slightly higher integrity, while the LTE-produced proteins have the lowest aggregation propensity among the systems analyzed. The generally low aggregation propensity and high integrity of HeLa- and LTE-produced proteins enable their analysis without purification. Since the LTE system is based on a fermentable organism with fast growth rates and cheap cultivation medium, it has potential cost advantages over the HeLa system. Furthermore, the genetic manipulation on *Leishmania tarentolae* is straightforward, allowing further improvements of the system.

This work was supported in part by ARC DP grant DP1094080, ARC FF FT0991611, LP120100162 and NHMRC project grant 569652 and NHMRC program grant APP1037320 to KA.

References

Aceti DJ, Bingman CA, Wrobel RL, Frederick RO, Makino S-I, Nichols KW, Sahu SC, Bergeman LF, Blommel PG, Cornilescu CC, Gromek KA, Seder KD, Hwang S, Primm JG, Sabat G, Vojtik FC, Volkman BF, Zolnai Z, Phillips GN, Markley JL, Fox BG. 2015. Expression platforms for producing eukaryotic proteins: A comparison of *E. coli* cell-based and wheat germ cell-free synthesis, affinity and solubility tags, and cloning strategies. J Struct Funct Genomics 16:67–80.

Ahn J-H, Keum J-W, Kim D-M. 2011. Expression screening of fusion partners from an *E. coli* genome for soluble expression of recombinant proteins in a cell-free protein synthesis system. PLoS ONE 6:e26875.

Barchinger SE, Ades SE. 2013. Regulated proteolysis: Control of the *Escherichia coli* σ (E)-dependent cell envelope stress response. Subcell Biochem 66:129–160.

Benyair R, Ron E, Lederkremer GZ. 2011. Protein quality control, retention, and degradation at the endoplasmic reticulum. Int Rev Cell Mol Biol 292:197–280.

Buntru M, Vogel S, Spiegel H, Schillberg S. 2014. Tobacco BY-2 cell-free lysate: An alternative and highly-productive plant-based in vitro translation system. BMC Biotechnol 14:37.

Chong S. 2014. Overview of cell-free protein synthesis: Historic landmarks, commercial systems, and expanding applications. Curr Protoc Mol Biol 108:16.30.1–16.30.11.

Endo Y, Sawasaki T. 2006. Cell-free expression systems for eukaryotic protein production. Curr Opin Biotechnol 17(4):373–380.

Fulda S, Gorman AM, Hori O, Samali A. 2010. Cellular stress responses: Cell survival and cell death. Int J Cell Biol 2010:214074.

Gagoski D, Mureev S, Giles N, Johnston W, Dahmer-Heath M, Škalamera D, Gonda TJ, Alexandrov K. 2015. Gateway-compatible vectors for high-throughput protein expression in pro- and eukaryotic cell-free systems. J Biotechnol 195:1–7.

Gambin Y, Ariotti N, McMahon K-A, Bastiani M, Sierecki E, Kovtun O, Polinkovsky ME, Magenau A, Jung W, Okano S, Zhou Y, Leneva N, Mureev S, Johnston W, Gaus K, Hancock JF, Collins BM, Alexandrov K, Parton RG. 2014. Single-molecule analysis reveals self assembly and nanoscale segregation of two distinct cavin subcomplexes on caveolae. Elife 3:e01434.

Hillebrecht JR, Chong S. 2008. A comparative study of protein synthesis in in vitro systems: From the prokaryotic reconstituted to the eukaryotic extract-based. BMC Biotechnol 8:58.

Hirano N, Sawasaki T, Tozawa Y, Endo Y, Takai K. 2006. Tolerance for random recombination of domains in prokaryotic and eukaryotic translation systems: Limited interdomain misfolding in a eukaryotic translation system. Proteins 64:343–354.

Ishihara G, Goto M, Saeki M, Ito K, Hori T, Kigawa T, Shirouzu M, Yokoyama S. 2005. Expression of G protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fusion vectors. Protein Expr Purif 41:27–37.

Johnston WA, Alexandrov K. 2014. Production of eukaryotic cell-free lysate from *Leishmania tarentolae*. Methods Mol Biol 1118:1–15.

Kovtun O, Mureev S, Jung W, Kubala MH, Johnston W, Alexandrov K. 2011. *Leishmania* cell-free protein expression system. Methods 55(1):58–64.

Kovtun O, Mureev S, Johnston W, Alexandrov K. 2010. Towards the construction of expressed proteomes using a *Leishmania tarentolae* based cell-free expression system. PLoS ONE 5(12):e14388.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187–208.

Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P, Rogers J, Lawlor S, McLaren S, Dricot A, Borick H, Cusick ME, Vandenhaute J, Dunham I, Hill DE, Vidal M. 2007. HORFeome v3.1: A resource of human open reading frames representing over 10,000 human genes. Genomics 89:307–315.

Martinou JC, Desagher S, Antonsson B. 2000. Cytochrome c release from mitochondria: All or nothing. Nat Cell Biol 2:E41–E43.

Matsuda T, Kigawa T, Koshiba S, Inoue M, Aoki M, Yamasaki K, Seki M, Shinozaki K, Yokoyama S. 2006. Cell-free synthesis of zinc-binding proteins. J Struct Funct Genomics 7:93–100.

Mureev S, Kovtun O, Nguyen UT, Alexandrov K. 2009. Species-independent translational leaders facilitate cell-free expression. Nat Biotechnol 27:747–752.

Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47:1588–1602.

Ramachandiran V, Kramer G, Hardesty B. 2000. Expression of different coding sequences in cell-free bacterial and eukaryotic systems indicates translational pausing on *Escherichia coli* ribosomes. FEBS Lett 482:185–188.

Rosenblum G, Cooperman BS. 2013. Engine out of the chassis: Cell-free protein synthesis and its uses. FEBS Lett 588(2):261–268.

Rual J-F, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain P-O, Clingingsmith TR, Hartley JL, Esposito D, Cheo D, Moore T, Simmons B, Sequerra R, Bosak S, Doucette-Stamm L, Le Peuch C, Vandenhaute J, Cusick ME, Albala JS, Hill DE, Vidal M. 2004. Human ORFeome version 1.1: A platform for reverse proteomics. Genome Res 14:2128–2135.

Saul J, Petritis B, Sau S, Rauf F, Gaskin M, Ober-Reynolds B, Mineyev I, Magee M, Chaput J, Qiu J, LaBaer J. 2014. Development of a full-length human protein production pipeline. Protein Sci 23:1123–1135.

Schwarz D, Junge F, Durst F, Frölich N, Schneider B, Reckel S, Sobhanifar S, Dötsch V, Bernhard F, Frölich N, Dötsch V. 2007. Preparative scale expression of membrane proteins in *Escherichia coli*-based continuos exchange cell-free systems. Nat Protoc 11:2945–2957.

Shimono K, Goto M, Kikukawa T, Miyauchi S, Shirouzu M, Kamo N, Yokoyama S. 2009. Production of functional bacteriorhodopsin by an *Escherichia coli* cell-free protein synthesis system supplemented with steroid detergent and lipid. Protein Sci 18:2160–2171.

Shin J, Noireaux V. 2010. Study of messenger RNA inactivation and protein degradation in an *Escherichia coli* cell-free expression system. J Biol Eng 4:9.

Škalamera D, Ranall MV, Wilson BM, Leo P, Purdon AS, Hyde C, Nourbakhsh E, Grimmond SM, Barry SC, Gabrielli B, Gonda TJ. 2011. A high-throughput platform for lentiviral overexpression screening of the human ORFeome. PLoS ONE 6(5):e20057.

Tan C, Saurabh S, Bruchez MP, Schwartz R, Leduc P. 2013. Molecular crowding shapes gene expression in synthetic cellular nanosystems. Nat Nanotechnol 8:602–608.

Tasaki T, Sriram SM, Park KS, Kwon YT. 2012. The N-end rule pathway. Annu Rev Biochem 81:261–289.

Terada T, Murata T, Shirouzu M, Yokoyama S. 2014. Cell-free expression of protein complexes for structural biology. Methods Mol Biol 1091:151–159.

Wang X, Zhao L, Zhao K-N. 2014. An optimized yeast cell-free lysate system for in vitro translation of human virus mRNA. Methods Mol Biol 1118:219–230.

Zawada JF, Yin G, Steiner AR, Yang J, Naresh A, Roy SM, Gold DS, Heinsohn HG, Murray CJ. 2011. Microscale to manufacturing scale-up of cell-free cytokine production-a new approach for shortening protein production development timelines. Biotechnol Bioeng 108:1570–1578.

Zhou Y, Asahara H, Gaucher EA, Chong S. 2012. Reconstitution of translation from Thermus thermophilus reveals a minimal set of components sufficient for protein synthesis at high temperatures and functional conservation of modern and ancient translation components. Nucleic Acids Res 40:7932–7945.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.