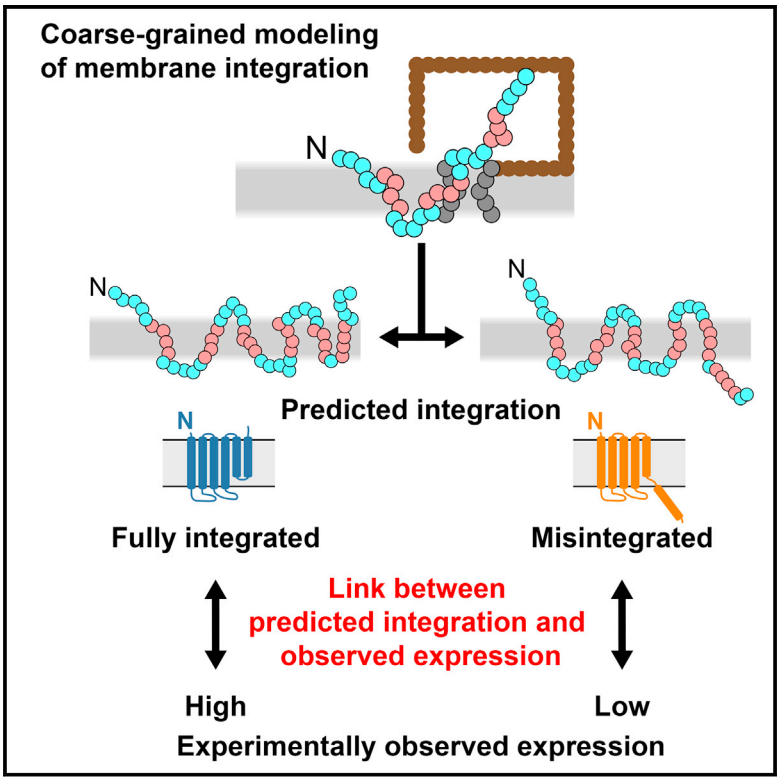
Article

Cell Reports

A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency

Graphical Abstract



Authors

Stephen S. Marshall, Michiel J.M. Niesen, Axel Müller, ..., Bin Zhang, William M. Clemons, Jr., Thomas F. Miller III

Correspondence

clemons@caltech.edu (W.M.C.), tfm@caltech.edu (T.F.M.)

In Brief

Marshall et al. demonstrate that an important bottleneck for integral membrane protein (IMP) expression is integration into the membrane. A recently developed computational model for predicting IMP integration efficiency is used to understand, predict, and enhance experimentally observed IMP expression levels.

Highlights

- Closely related IMP sequences exhibit substantial differences in expression levels
- IMP expression levels are dependent upon the efficiency of membrane integration
- In distinct systems, mutations that improve IMP integration also improve expression
- Simulated IMP integration is used to design IMPs with enhanced expression

Marshall et al., 2016, Cell Reports *16*, 1–9
August 23, 2016
http://dx.doi.org/10.1016/j.celrep.2016.07.042

CellPress

Please cite this article in press as: Marshall et al., A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency, Cell Reports (2016), http://dx.doi.org/10.1016/j.celrep.2016.07.042

Cell Reports
Article

A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency

Stephen S. Marshall,[1,2] Michiel J.M. Niesen,[1,2] Axel Müller,[1] Katrin Tiemann,[1] Shyam M. Saladi,[1] Rachel P. Galimidi,[1] Bin Zhang,[1] William M. Clemons, Jr.,[1,3,*] and Thomas F. Miller III[1,*]
[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Co-first author
[3]Lead Contact
*Correspondence: clemons@caltech.edu (W.M.C.), tfm@caltech.edu (T.F.M.)
http://dx.doi.org/10.1016/j.celrep.2016.07.042

SUMMARY

Integral membrane proteins (IMPs) control the flow of information and nutrients across cell membranes, yet IMP mechanistic studies are hindered by difficulties in expression. We investigate this issue by addressing the connection between IMP sequence and observed expression levels. For homologs of the IMP TatC, observed expression levels vary widely and are affected by small changes in protein sequence. The effect of sequence changes on experimentally observed expression levels strongly correlates with the simulated integration efficiency obtained from coarse-grained modeling, which is directly confirmed using an in vivo assay. Furthermore, mutations that improve the simulated integration efficiency likewise increase the experimentally observed expression levels. Demonstration of these trends in both *Escherichia coli* and *Mycobacterium smegmatis* suggests that the results are general to other expression systems. This work suggests that IMP integration is a determinant for successful expression, raising the possibility of controlling IMP expression via rational design.

INTRODUCTION

The central role of integral membrane proteins (IMPs) in many biological functions motivates structural and biophysical studies that require large amounts of purified protein, often at considerable costs in terms of both materials and labor. A key obstacle is that only a small percentage of IMPs can be overexpressed (i.e., heterologously produced at levels conducive to further study) (Lewinson et al., 2008). While extensive efforts have shown promising results for individual IMPs, including those focusing on expression conditions, host modification, and directed evolution (reviewed in Schlegel et al., 2010; Wagner et al., 2006; and Scott et al., 2013), none of these has proven broadly applicable, even among homologs of a given IMP. In general, the determinants for IMP expression are poorly understood, leading to the prevailing opinion that problems in membrane protein expression must be addressed on a case-by-case basis.

Closely related IMP homologs can vary dramatically in the amount of protein available after expression (Lewinson et al., 2008), which raises a fundamental question: what differentiates the expression of IMP homologs? The hypothesis raised here is that the efficiency with which an IMP is integrated into the membrane is a key determinant in the degree of observed IMP expression.

A fundamental step in the biosynthesis of most IMPs involves their targeting to and integration into the membrane via the Sec protein translocation channel (Rapoport, 2007). Integration of IMP transmembrane domains (TMDs) into the membrane is facilitated primarily through interaction between the nascent chain and SecY, which forms the core of the protein translocation complex, or translocon. Following the co-translational or post-translational insertion of nascent protein sequences into the translocon channel, hydrophobic segments pass through the lateral gate of SecY into the membrane to form TMDs. Factors such as TMD hydrophobicity (Harley et al., 1998; Hessa et al., 2005) and loop charge (Heijne, 1986; Goder and Spiess, 2003) have been shown to affect the efficiency of TMD integration and topogenesis. For example, TMD hydrophobicity is directly related to the probability with which TMDs partition into the lipid bilayer, while positively charged residues in the loop alter TMD orientation by preferentially occupying the cytosol (Goder and Spiess, 2003; Hessa et al., 2005; Heijne, 1986).

In this study, we investigated the connection between observed IMP expression levels and Sec-facilitated IMP integration efficiency (i.e., the probability of membrane integration with the correct multi-spanning topology). Systematic investigation of chimeras within an IMP family led to the identification of sequence elements that modulate expression levels. In silico modeling of IMP integration at the Sec translocation channel found that the sequence modifications that increase the calculated IMP integration efficiency correlate with in vivo overexpression improvements, suggesting that IMP integration efficiency is a determinant for successful expression. The result was found to be general across distinct expression systems (*E. coli* and *M. smegmatis*). Furthermore, an in vivo assay based on antibiotic resistance in *E. coli* experimentally confirmed the model that the integration efficiency of an individual TMD correlates with the observed IMP expression levels. The strong link between

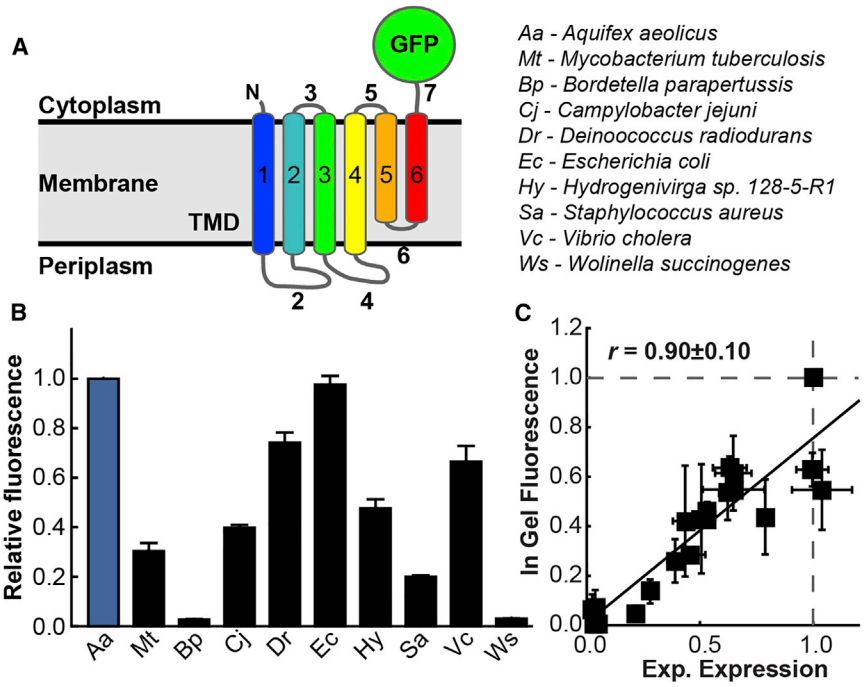


Figure 1. Variation in the Expression of TatC Homologs in *E. coli*

(A) A topology representation of TatC with a GFP C-terminal tag, as used in the expression studies. TMDs and loops are indicated in colors and gray, respectively, and are numbered.

(B) Expression levels of various TatC homologs in *E. coli*, measured by TatC-GFP fluorescence, with expression levels normalized to *Aa*TatC (blue). Error bars indicate the SEM.

(C) Correlation of the in-gel fluorescence quantified for each band versus the experimental expression measured by flow cytometry. Both metrics are highly correlated across multiple trials (r is the Pearson correlation coefficient), with in-gel fluorescence showing the same trends in expression yield as seen by flow-cytometry. Error bars indicate the SEM. See also Figure S1.

the effect of sequence modifications on simulated integration efficiency and experimentally measured expression levels offers future promise for the rational design of IMP systems with increased expression levels.

RESULTS

As a detailed case study, the TatC IMP family was employed for all experimental and computational results reported here. A component of the bacterial twin-arginine translocation pathway, TatC plays a key role in the transport of folded proteins across the cytoplasmic membrane (Bogsch et al., 1998). The employment of TatC was well suited for this study as it is reasonably sized (only six TMDs; Figure 1A), non-essential, and found broadly throughout bacteria; furthermore, TatC homologs previously have been observed to exhibit widely varying expression levels in *E. coli* (Ramasamy et al., 2013), suggesting the importance of sequence-level details in the expression of this IMP.

Wild-Type and Chimeric TatC Expression in *E. coli*

We first demonstrated that homologs of the IMP TatC exhibit large variance in observed expression levels in *E. coli*. For a quantitative measure of IMP expression, we employed a C-terminal fusion tag of a GFP variant (Waldo et al., 1999) (Figure 1A) and measured whole-cell fluorescence by flow cytometry. Whole-cell fluorescence intensity of this fusion tag has been validated in numerous previous studies to correlate strongly with the amount of folded IMP, rather than the total level of IMP translated (Fluman et al., 2014; Wang et al., 2011; Guglielmi et al., 2011; Geertsma et al., 2008; Drew et al., 2005). We further validated the expression levels measured from whole-cell fluorescence (Figure 1B) using in-gel fluorescence (Figure 1C; Figure S1; Pearson correlation coefficient, r = 0.9) and western blot analysis (Figure S1). With this approach, expression levels in *E. coli* were experimentally measured for TatC homologs from a variety of bacteria, including *Aquifex aeolicus* (*Aa*), *Bordetella parapertussis* (*Bp*), *Campylobacter jejuni* (*Cj*), *Deinococcus radiodurans* (*Dr*), *Escherichia coli* (*Ec*), *Hydrogenivirga* species 128-5-R1 (*Hy*), *Mycobacterium tuberculosis* (*Mt*), *Staphylococcus aureus* (*Sa*), *Vibrio cholera* (*Vc*), and *Wolinella succinogenes* (*Ws*) (sequences in Figure S2).

Figure 1B shows the wide range of expression levels that are exhibited by the TatC homologs in *E. coli*. Previous expression trials of TatC homologs identified that *Aa*TatC is readily produced at high levels in *E. coli*, which enabled the solution of its structure (Ramasamy et al., 2013; Rollauer et al., 2012). In contrast, low expression is found for both the *Mt*TatC, hereafter referred to as *Mt*TatC(Wt-tail), and a modified sequence truncating the un-conserved 38-residue sequence of the C-terminal loop, hereafter referred to as *Mt*TatC (Ramasamy et al., 2013).

To examine the parts of the protein sequence that affect expression, swap chimeras were generated by exchanging entire loops and TMDs between *Aa*TatC and *Mt*TatC (sequences in Table S1). The TMDs and loops were defined by comparing sequence alignments and membrane topology predictions (Figure 2B) (Sievers et al., 2011; Tsirigos et al., 2015). The swap chimeras exhibited a wide range of expression results (Figure 2A). The C-terminal loop sequence, referred to as the C-tail and labeled as loop 7 in Figure 1A, was found to have a significant effect on expression levels (shaded bars in Figure 2A). Removal of the *Mt*TatC C-tail improved expression. Removal of the C-tail from the *Aa*TatC sequence led to a corresponding decrease in expression. Strikingly, swapping the *Aa*TatC C-tail (*Aa*-tail) into the *Mt*TatC sequence led to a significant improvement in expression.

The positive effect of the *Aa*-tail on *Mt*TatC expression raises the question of whether expression can be similarly improved in other TatC homologs by substituting the corresponding C-tail sequence (Figure 2E) with that of *Aa*TatC. Swapping the C-tail

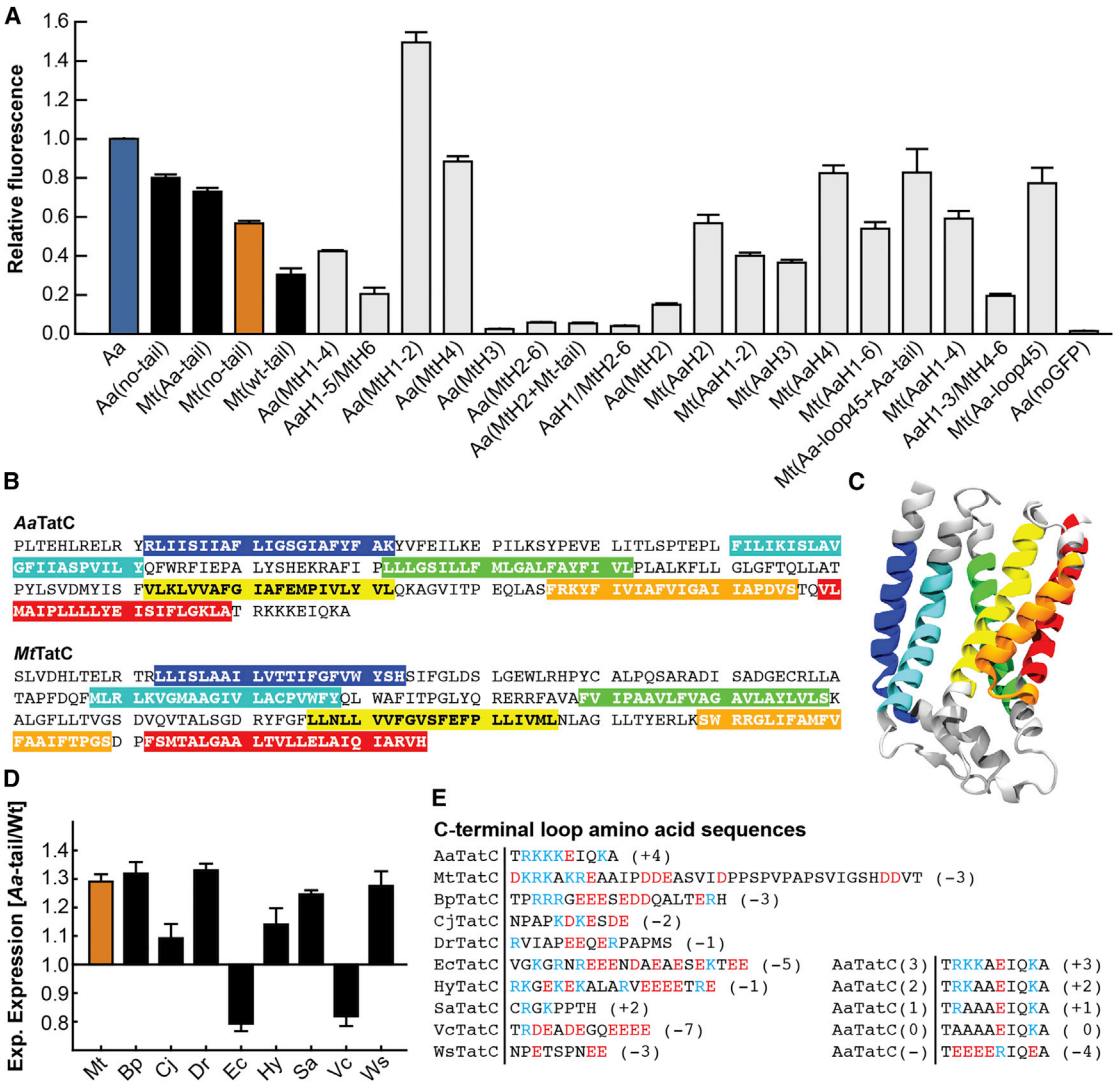


Figure 2. Effect of the C-tail on TatC Expression in *E. coli*

(A) Measured expression levels of the *Aa*TatC and *Mt*TatC chimera proteins, normalized to *Aa*TatC. Shaded bars represent wild-type TatC homologs and mutants with C-tail modifications.

(B) Domain definitions used in generating the swap chimeras, with TMDs highlighted, are shown.

(C) A ribbons diagram of the structure of *Aa*TatC (RCSB PDB: 4HTS). TMDs are colored according to the highlights used in (B).

(D) For each homolog, the ratio of the measured expression level for the *Aa*-tail chimera to that of the corresponding wild-type sequence is shown.

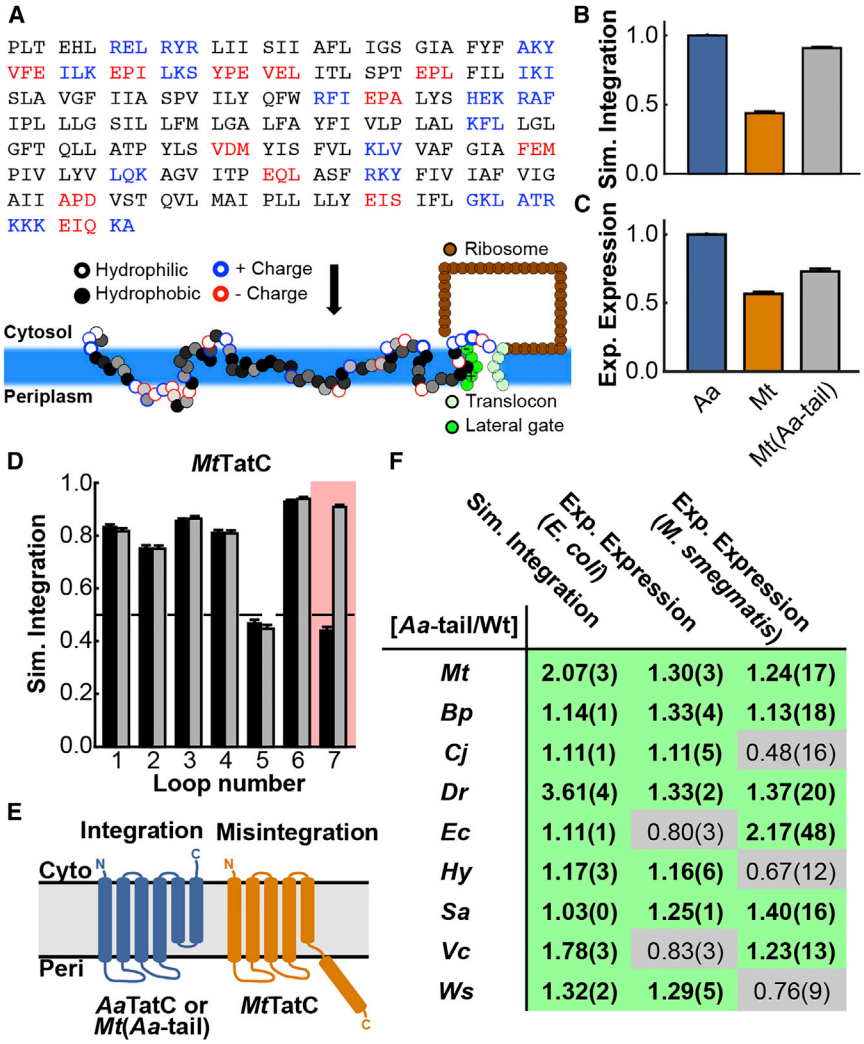
(E) TatC wild-type and charge mutant C-tail sequences. Positive residues are in blue and negative residues are in red. The net charge is shown to the right of each sequence.

Error bars indicate the SEM.

of the various TatC homologs with the *Aa*-tail improved expression in seven of nine cases (Figure 2D). Taken together, the results in Figure 2 indicate that the C-tail is a significant factor in determining TatC expression across homologs.

In Silico Modeling of TatC Integration

To investigate the mechanistic basis for the experimentally observed effect of the C-tail on expression, we employed a recently developed in silico coarse-grained (CG) approach that models co-translational translocation on unbiased biological timescales (Zhang and Miller, 2012b). The CG model, which is derived from >16 μs of molecular dynamics simulations of the Sec translocation channel, the membrane bilayer, and protein substrates (Zhang and Miller, 2010, 2012a), has been validated for the description of Sec-facilitated membrane integration, including experimentally observed effects of amino acid sequence on the membrane topology of single-spanning IMPs (Zhang and Miller, 2012b) and multi-spanning dual-topology proteins (Van Lehn et al., 2015). IMP sequences were mapped onto a Brownian dynamics model of the ribosome/translocation channel/nascent protein system, and the Sec translocon-facilitated integration of the IMP into the lipid bilayer was directly



F

[*Aa*-tail/Wt]	Sim. Integration	Exp. Expression (*E. coli*)	Exp. Expression (*M. smegmatis*)
Mt	**2.07(3)**	**1.30(3)**	**1.24(17)**
Bp	**1.14(1)**	**1.33(4)**	**1.13(18)**
Cj	**1.11(1)**	**1.11(5)**	0.48(16)
Dr	**3.61(4)**	**1.33(2)**	**1.37(20)**
Ec	**1.11(1)**	0.80(3)	**2.17(48)**
Hy	**1.17(3)**	**1.16(6)**	0.67(12)
Sa	**1.03(0)**	**1.25(1)**	**1.40(16)**
Vc	**1.78(3)**	0.83(3)	**1.23(13)**
Ws	**1.32(2)**	**1.29(5)**	0.76(9)

Figure 3. Calculation of TatC Integration Efficiencies

(A) Schematic illustration of the CG simulation model that is used to model co-translational IMP membrane integration. The amino acid sequence of the IMP is mapped onto CG beads, with each consecutive trio of amino acid residues in the nascent protein sequence mapped to an associated CG bead; the underlying properties of the amino acid residues determine the interactions of the CG beads, as described in the text.

(B) Simulated integration efficiency of the *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail) sequences is shown. Error bars indicate the SEM.

(C) Experimental expression of the *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail) sequences is shown. Error bars indicate the SEM.

(D) The simulated integration efficiency for individual loops of both the wild-type *Mt*TatC sequence (black bars) and the *Aa*-tail swap chimera (gray bars), with loop 7 highlighted, is shown. Error bars indicate the SEM.

(E) Schematic of the correct and incorrect TatC topologies observed in the simulations. Misintegration of loop 7 and translocation of TMD 6 lead to an incorrect final topology for *Mt*TatC.

(F) For each homolog, comparison between the experimental expression levels in *E. coli* and *M. smegmatis* and the simulated integration efficiencies, reporting the ratio of the *Aa*-tail chimera result to that of the corresponding wild-type sequence. Ratios exceeding unity are highlighted in green, indicating enhancement due to the *Aa*-tail. Values in parentheses indicate the SEM. See also Figure S4.

simulated in 1,200 independent minute-timescale trajectories for each TatC (Figure 3A). This implementation of the CG model did not distinguish between expression systems.

Using the results of the CG model, Figure 3B presents the simulated integration efficiency (i.e., the simulated integration efficiency is defined to be the fraction of trajectories that led to the correct membrane topology) for several TatC sequences. Unless otherwise specified, we defined membrane topology in terms of the final orientation of the C-tail; Figure S3 confirms that analyzing the trajectories in terms of this single-loop definition for membrane topology correlates with defining topology in terms of all loops, while reducing the statistical noise. The *Aa*TatC homolog exhibited significantly higher simulated integration efficiency than the *Mt*TatC homolog, which is consistent with the relative experimental expression levels for the two homologs in Figure 3C. Figure 3B shows that the *Mt*(*Aa*-tail) chimera recovered the high levels of simulated integration efficiency seen for the *Aa*TatC homolog, further mirroring the experimental trends in IMP expression (Figure 3C). Figure 3D presents an analysis of the orientation of each loop, indicating that only loop 7 was significantly affected swapping the C-tail in the simulations. As is shown schematically in Figure 3E, the simulations found that *Mt*TatC exhibits a large fraction of trajectories in which the C-tail resides in the periplasm, such that the C-terminal TMD (TMD 6) fails to correctly integrate into the membrane.

Additional simulations were performed for the full set of the experimentally characterized TatC homologs (Figure S4), allowing comparison of the computationally predicted shifts in IMP integration with those observed experimentally for IMP expression. For each homolog, Figure 3F compares the effect of swapping the wild-type C-tail with the *Aa*-tail on both the experimental expression level and the simulated integration efficiency. With the exception of *Vc*TatC and *Ec*TatC, Figure 3F shows consistent agreement between the computational and experimental results in *E. coli* upon introducing the *Aa*-tail.

Confirmation of the Predicted Mechanism

The comparison between simulation and experiment in the previous sections suggests a mechanism in which translocation of the C-tail of TatC into the periplasm leads to a reduction in the observed expression level. To validate this, an experimental in vivo assay based on antibiotic resistance in *E. coli* was employed. The C-terminal GFP tag was replaced by β-lactamase,

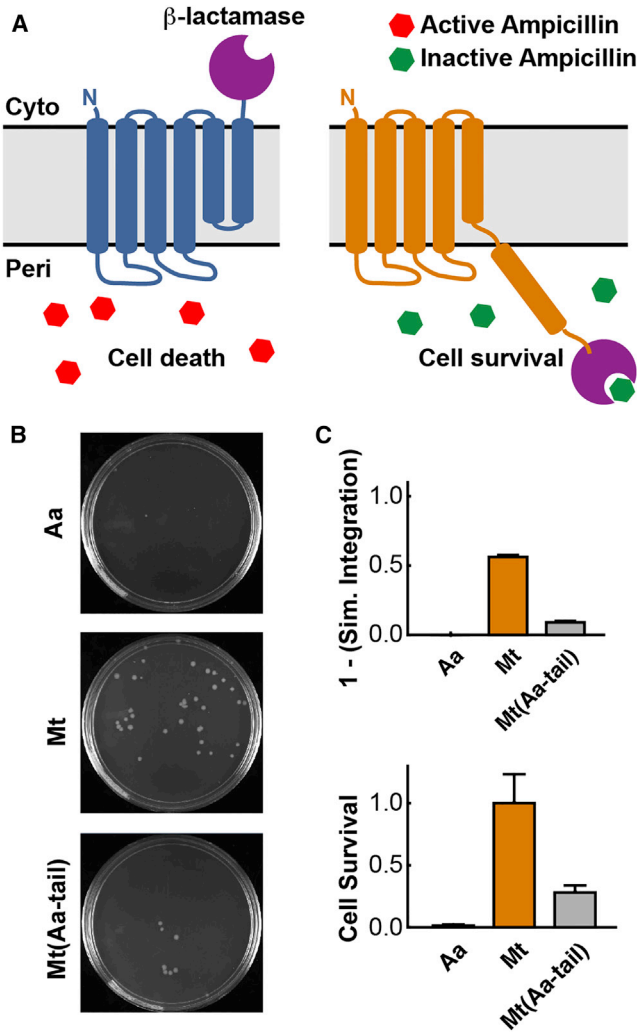


Figure 4. Correlation of Antibiotic Resistance to Membrane Topology

(A) Schematic of the cytoplasmic and periplasmic topologies of the TatC C-tail with the fused β-lactamase enzyme. Misintegration of loop 7 leads to periplasmic localization of the β-lactamase, resulting in enhanced antibiotic resistance and cell survival.

(B) Representative plates from the ampicillin survival test are shown.

(C) Comparison of the simulated integration efficiency (top) and relative ampicillin survival rate (bottom) for *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail). The reported cell survival corresponds to the ratio of counted cells post-treatment versus prior to treatment with ampicillin; all values are reported relative to *Mt*TatC. Error bars indicate the SEM.

such that an incorrectly oriented C-tail would confer increased resistance to β-lactam antibiotics (Figure 4A); an inverse correlation between antibiotic resistance and GFP fluorescence was thus expected. *Aa*TatC, *Mt*, and *Mt*(*Aa*-tail) constructs containing the β-lactamase tag were expressed using the same protocol as before. Following expression, the cells were diluted to an optical density 600 (OD_{600}) of 0.1 in fresh media without inducing agent, and they were grown to an OD_{600} of ~0.5 at which point ampicillin was added. Then 1.5 hr after ampicillin treatment, equal amounts of the media were plated on Luria-Bertani (LB) agar plates without ampicillin (Figure 4B). The number of observed colonies was used to quantify the relative cell survival (Figure 4C, bottom). The survival rate of *Mt*(*Aa*-tail), *Mt*, and *Aa*TatC inversely correlated with the simulated integration efficiency of the C-tail (Figure 4C), validating the proposed mechanism.

Tail Charge as an Expression Determinant: Experimental Tests of Computational Predictions

To further establish the connection between the simulated integration efficiencies and the experimentally observed expression levels, we examined the effect of C-tail mutations. We focused on modifications of the C-tail amino acid sequences that involve the introduction or removal of charged residues, which are known to affect IMP topology and stop-transfer efficiency (Goder and Spiess, 2003; Seppälä et al., 2010; Zhang and Miller, 2012b).

We began by investigating the generic effect of the C-tail charge magnitude on TatC-simulated integration efficiency. Figure 5A presents the results of CG simulations in which the magnitude of the charges on the C-tail of the *Mt*(*Aa*-tail) sequence were scaled by a multiplicative factor, χ, keeping all other aspects of the protein sequence unchanged. The simulations revealed that reducing the charge magnitude on the C-tail led to lower simulated integration efficiency.

To examine the corresponding effect of C-tail charge magnitude on expression levels, Figure 5B plots the ratio of experimentally observed expression for each wild-type homolog relative to its corresponding *Aa*-tail swap chimera versus the total charge magnitude on the wild-type C-tail. Without exception in these data, the expression of wild-type homologs with weakly charged C-tails (relative to the *Aa*-tail) was improved upon swapping with the *Aa*-tail, whereas the expression of homologs with strongly charged C-tails was reduced upon swapping with the *Aa*-tail (i.e., all data points in Figure 5B fall into the unshaded quadrants).

Figure 5C further illustrates the effect of charge magnitude on expression by presenting the experimentally observed expression levels for *Aa*-tail(−) swap chimeras, in which the introduced C-tail sequence preserved the charge magnitude of the *Aa*-tail sequence while reversing the net charge (see Figure 2E for the C-tail sequences). Despite the complete reversal of the C-tail charge, the observed correlation between expression and C-tail charge magnitude for these two sets of chimeras was strikingly similar (compare Figures 5B and 5C).

Finally, we considered a series of mutants of the *Mt*(*Aa*-tail) chimera, in which the charge magnitude of the *Aa*-tail was reduced by mutating positively charged residues to alanine residues (see Figure 2E for the C-tail sequences). For this series of mutants, Figure 5D (black) shows that the simulated integration efficiency decreased with the charge of the C-tail, which predicted a corresponding decrease in the experimental expression levels; indeed, the subsequent experimental measurements confirmed the predicted trend (Figure 5D, blue). Again using the antibiotic resistance assay to validate the connection between simulated integration efficiency and observed expression, Figure 5E confirms that the simulation results correlated with the relative survival of the *Mt*(*Aa*-tail) alanine mutants with a β-lactamase tag (Figure 5E, red). In addition to providing evidence for

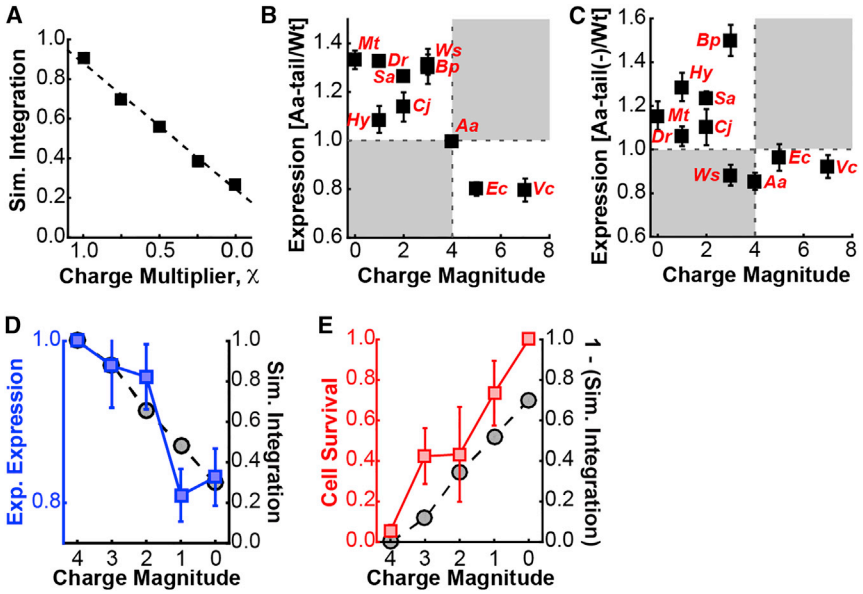


Figure 5. Mechanistic Basis Associated with Charged C-tail Residues

(A) Simulated integration efficiency of the *Mt*(*Aa*-tail) chimera, as a function of scaling the charges of the C-tail residues, is shown.

(B) Correlation of the ratio of the measured expression for the *Aa*-tail swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail (data from Figures 2B and 2E). (Pearson correlation coefficient of r = 0.8 ± 0.2)

(C) Correlation of the ratio of the measured expression for the *Aa*-tail(−) swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail, where the *Aa*-tail(−) swap chimeras include a variant of the *Aa*-tail with net negative charge and the same overall charge magnitude, is shown.

(D) Experimental expression levels in *E. coli* (blue, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Reported values are normalized to *Mt*(*Aa*-tail).

(E) Relative ampicillin survival rate in *E. coli* (red, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Simulation results are normalized as in (D), while ampicillin survival is normalized to the highest survival rate (i.e., with zero charge magnitude). Error bars indicate the SEM.

the connection between simulated integration efficiency and observed expression levels, the results in Figure 5 suggest that this link can be used to control IMP expression.

Transferability to Another Expression System

Beyond the *E. coli* overexpression host, we examined the transferability of the relation between simulated integration efficiency and experimental expression levels. We employed *M. smegmatis*, a genetically tractable model organism that is phylogenetically distinct from *E. coli*. All coding sequences were transferred into an inducible *M. smegmatis* vector, including the linker and C-terminal GFP, and expressed; expression levels were then measured by flow cytometry and validated by western blot.

Figure 6A shows that, as in *E. coli*, the experimentally observed expression levels vary widely among the wild-type TatC homologs in *M. smegmatis*. However, comparison of Figure 6A with Figure 1B reveals that the total expression levels for the homologs in *M. smegmatis* are different from those seen in *E. coli*, although for both systems the *Aa*TatC homolog expresses strongly and *Mt*TatC expresses poorly (which is perhaps surprising, given the close evolutionary link between *M. smegmatis* and *M. tuberculosis*). Figure 3F also shows that replacing the wild-type C-tail with the *Aa*-tail in *M. smegmatis* generally increased the experimentally observed expression levels, in general agreement (six of nine homologs) with the previously discussed simulated integration efficiency results.

Figure 3F further shows that the subset of homologs, for which the *Aa*-tail swap chimeras led to increased levels of expression in *M. smegmatis*, was overlapping but different from the subset associated with the *E. coli* results. This emphasizes that, although the computed levels of simulated integration efficiency agree with the observed changes in expression levels in both expression systems, the observed expression levels depend on the expression system, while the simulated integration efficiencies calculated using the current implementation of the CG model are independent of the expression system. In short, simulated integration efficiency is a predictor of the expression levels in both systems, but it is not the only factor contributing to the observed expression levels.

Continuing with the *M. smegmatis* expression system, Figure 6B repeats the comparison between the simulated integration efficiency and the observed expression levels for the series of mutants of the *Mt*(*Aa*-tail) chimera, in which the positive charge of the *Aa*-tail was reduced by mutating positively charged residues to alanine residues. The simulated integration efficiencies, identical to those in Figure 5D, were predicted to decrease as charges were removed. The experimental expression levels for *M. smegmatis* in Figure 6B likewise showed a decrease. Taken together, the results obtained for the *M. smegmatis* expression system suggest that the connection between simulated integration efficiency and observed expression levels may be generalizable beyond *E. coli.*

Transferability beyond the C-tail: Analysis of Loop 5 Swap Chimera

As seen in Figure 3D, the CG simulations predicted poor integration efficiency for loop 5, suggesting an additional location (beyond the C-tail, loop 7) in the *Mt*TatC sequence that could be optimized for expression. Figure 7A presents the simulated integration efficiency for loop 5 in each of the TatC homologs, revealing a significant range of efficiencies. Selecting the four homologs with the highest predicted simulated integration efficiency for loop 5 (*Sa*, *Hy*, *Cj*, and *Vc*), chimera proteins were

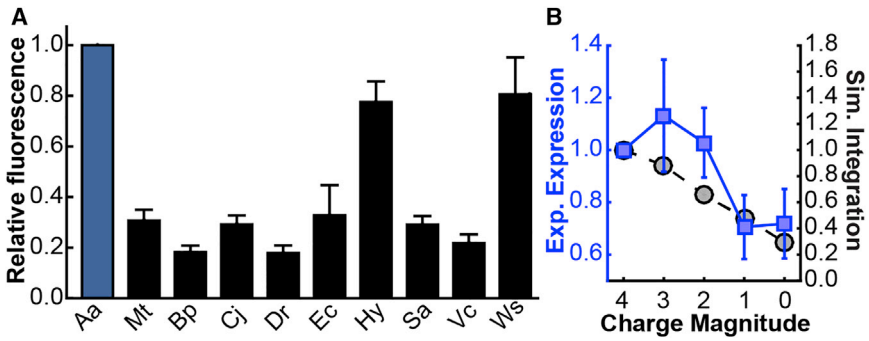


Figure 6. *M. smegmatis* Expression Tests

(A) Expression levels of various TatC homologs in *M. smegmatis* were measured by TatC-GFP fluorescence, with expression levels normalized to *Aa*TatC (blue).

(B) Simulated integration efficiency (blue, left axis) and measured expression levels in *M. smegmatis* (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Error bars indicate the SEM.

derived from the *Mt*TatC sequence by swapping loop 5 of *Mt*TatC with the corresponding loop 5 sequence from each of these homologs (Figure 7B). Figure 7C compares the simulated integration efficiency and experimentally observed expression level for each chimera, revealing agreement for three of four cases. Comparing the simulation results in Figure 7, note that the degree of improvement for the simulated integration efficiency obtained from the CG simulations of the chimeras (Figure 7C) is different from that anticipated by naive comparison of the individual loops in the wild-type sequences (Figure 7A); this emphasizes that the simulated integration efficiency is sensitive to elements of the IMP sequence beyond the local segment that is being swapped. The results in Figure 7 suggest the simulated integration efficiency can be used to identify regions beyond the TatC C-tail for modification to improve experimental expression; more generally, they suggest the potential for identifying local segments of an IMP amino acid sequence that may be modified to yield increased experimental expression.

DISCUSSION

The mechanistic picture that emerges from the experimental and theoretical analysis of the TatC IMP family is that the efficiency of Sec-facilitated membrane integration, which is impacted by the IMP amino acid sequence, is a key determinant in the degree of observed protein expression. We observed that TatC homologs had varying levels of expression (Figures 1B and 6A). Swap chimeras between *Aa*TatC and *Mt*TatC revealed a significant effect of the C-tail in determining expression yields (Figure 2A), with the *Aa*-tail having a largely positive effect that was transferrable to other homologs (Figure 3F). CG modeling predicted a large, sequence-dependent variation of the simulated integration efficiency for the C-tail (Figure 3), suggesting the underlying mechanism by which the *Aa*-tail enhances the expression of other TatC homologs. Validation of this mechanism was experimentally demonstrated using an antibiotic resistance assay (Figure 4). Additional point-charge mutations in the C-tail were shown to change the simulated integration efficiency, which in turn predicted changes in the IMP expression levels according to the proposed mechanism; these predictions were experimentally confirmed in both *E. coli* (Figure 5) and *M. smegmatis* (Figure 6). Finally, the link between simulated integration efficiency and experimental expression was exploited to design *Mt*TatC chimeras with improved expression based on the loop 5 simulated integration efficiency (Figure 7).

The observed correlation between IMP integration efficiency and observed expression levels presented here is consistent with earlier observations that expression can be modulated by mutations of the sequence (Sarkar et al., 2008; Grisshammer et al., 1993; Warne et al., 2008), as well as recent work in which misintegrated dual-topology IMPs were shown to be degraded by FtsH (Woodall et al., 2015). However, these earlier studies did not provide a clear mechanistic basis for the relation between IMP sequence modifications and observed expression levels. In the current work, we demonstrate the relation between integration efficiency and observed expression levels, and we demonstrate a tractable CG approach for computing the simulated integration efficiency and its changes upon sequence modifications. This work also raises the possibility of using simulated integration efficiencies to optimize experimental expression levels, which has been demonstrated here via the computational prediction and subsequent experimental validation of individual charge mutations in the C-tail and of loop 5 swap chimeras.

A few comments are worthwhile with regard to the scope of the conclusions drawn here. First, our study focused on comparing protein expression levels among IMP sequences that involve relatively localized changes, such as single mutations or loop swap chimeras, as opposed to predicting relative expression levels among dramatically different IMP sequences. Second, our study examined experimental conditions for the overexpression of IMPs using the same plasmids, which may be expected to isolate the role of membrane integration in determining the relative expression levels of closely related IMP sequences. The prediction of expression levels among IMPs that involve more dramatic differences in sequence may well require the consideration of other factors, beyond just the simulated integration efficiency. Moving forward, we expect that a useful strategy will be to systematically combine the simulated IMP integration efficiency with other sequence-based properties to predict IMP expression levels (Daley et al., 2005).

The experimental and computational tools used here are readily applicable to many systems, potentially aiding the understanding and enhancement of IMP expression in many other systems, as well as providing fundamental tools for the investigation of co-translational IMP folding. By demonstrating inexpensive in silico methods for predicting protein expression, we note the potential for computationally guided protein

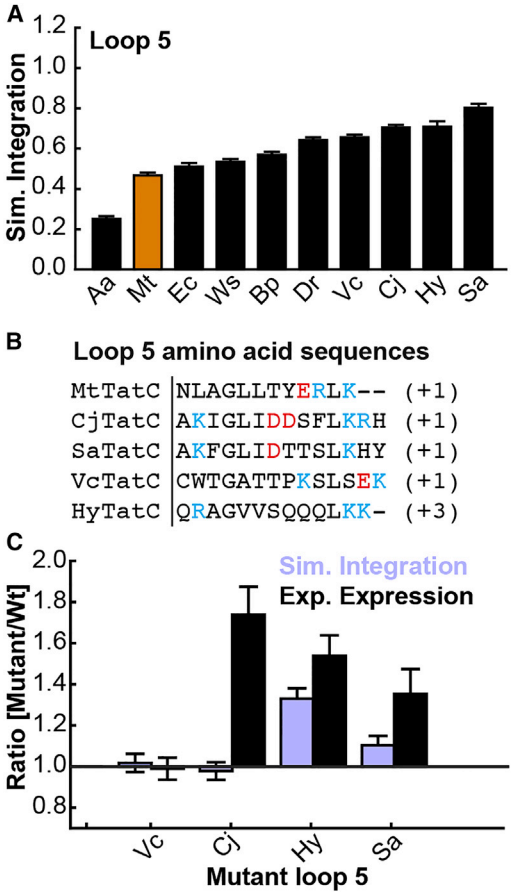


Figure 7. Loop 5 Analysis for *Mt*TatC

(A) Simulated integration efficiency of loop 5 for the TatC homologs is shown.
(B) Loop 5 amino acid sequence for various TatC homologs is shown.
(C) Experimental expression (black) and simulated integration efficiency (purple) for the loop 5 swap chimeras of *Mt*TatC, in which the entire loop 5 sequence of wild-type *Mt*TatC is replaced with the corresponding sequence of other homologs. Error bars indicate the SEM.

expression strategies to significantly impact the isolation and characterization of many IMPs.

EXPERIMENTAL PROCEDURES

Cloning, Expression, and Flow Cytometry

Briefly, for *E. coli* all expression plasmids were derived from pET28(a+)-GFP-ccdB, with the final expressed sequences containing a Met-Gly N terminus followed by the IMP sequence, a tobacco etch virus (TEV) protease site, a GFP variant (Waldo et al., 1999), and an eight His tag. For β-lactamase constructs, the GFP sequence was replaced by a β-lactamase sequence. For *M. smegmatis* expression plasmids, the entire coding region of the TatC homologs were sub-cloned and transferred into pMyNT vector (Noens et al., 2011). *E. coli* constructs were grown in BL21 Gold (DE3) cells (Agilent Technologies) at 16°C, induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at an OD_{600} of 0.3, and then analyzed after 16 hr. *M. smegmatis* constructs were grown in mc^2155 cells (ATCC) at 37°C, induced at an OD_{600} of 0.5 with 0.2% acetamide, and then analyzed after 6 hr. A 200-μl sample of each expression culture was pelleted and resuspended in 1 ml PBS. Whole-cell GFP fluorescence was measured using a MACSQuant10 Analyzer (Miltenyi Biotec). For the ampicillin survival assay, the cells were diluted to an OD_{600} of 0.1 in fresh media following expression without inducing agent and then grown to an OD_{600} of ~0.5, at which point ampicillin was added. Then 1.5 hr after ampicillin treatment, equal amounts of the media were plated on LB agar plates without ampicillin. The number of observed colonies was used to quantify the relative cell survival. Full experimental protocols are provided in the Supplemental Experimental Procedures.

Description of the CG Model

Modeling of IMP integration in the current study was performed using a previously developed CG method for the direct simulation of co-translational protein translocation and membrane integration (Zhang and Miller, 2012b). Ribosomal translation and membrane integration of nascent proteins are thus simulated on the minute timescale, enabling direct comparison between theory and experiment. The CG model previously was parameterized using extensive molecular dynamics simulations of the translocon and nascent protein in explicit lipid and water environments (Zhang and Miller, 2010, 2012a). The CG model has been validated against available experimental data, and it has been shown to correctly capture effects related to nascent protein charge, hydrophobicity, length, and translation rate in both IMP integration and protein translocation studies (Zhang and Miller, 2012b; Van Lehn et al., 2015). Details of the implementation of the CG model and the analysis of the simulated trajectories are given in the Supplemental Experimental Procedures and Table S2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.07.042.

AUTHOR CONTRIBUTIONS

S.S.M., M.J.M.N., A.M., W.M.C., and T.F.M. designed research. S.S.M. and M.J.M.N. performed research. S.S.M., M.J.M.N., A.M., K.T., S.M.S., R.P.G., and B.Z. contributed new reagents/analytic tools. S.S.M., M.J.M.N., A.M., W.M.C., and T.F.M. analyzed data. S.S.M., M.J.M.N., W.M.C., and T.F.M. wrote the paper.

ACKNOWLEDGMENTS

The authors thank R.C. Van Lehn and D.C. Rees for comments on the manuscript and D. Daley for helpful discussion of Daley et al. (2005). Work in the W.M.C. lab is supported by an NIH Pioneer Award to W.M.C. (5DP1GM105385) and an NIH training grant to S.S.M. (NIH/National Research Service Award (NRSA) training grant 5T32GM07616). Work in the T.F.M. group is supported in part by the Office of Naval Research (N00014-10-1-0884), and computational resources were provided by the National Energy Research Scientific Computing Center (NERSC), a (Department of Energy) DOE Office of Science User Facility (DE-AC02-05CH11231).

Received: January 4, 2016
Revised: June 9, 2016
Accepted: July 16, 2016
Published: August 11, 2016

REFERENCES

Bogsch, E.G., Sargent, F., Stanley, N.R., Berks, B.C., Robinson, C., and Palmer, T. (1998). An essential component of a novel bacterial protein export system with homologues in plastids and mitochondria. J. Biol. Chem. *273*, 18003–18006.

Daley, D.O., Rapp, M., Granseth, E., Melén, K., Drew, D., and von Heijne, G. (2005). Global topology analysis of the Escherichia coli inner membrane proteome. Science *308*, 1321–1323.

Drew, D., Slotboom, D.J., Friso, G., Reda, T., Genevaux, P., Rapp, M., Meindl-Beinker, N.M., Lambert, W., Lerch, M., Daley, D.O., et al. (2005). A scalable,

GFP-based pipeline for membrane protein overexpression screening and purification. Protein Sci. *14*, 2011–2017.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. (2014). mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. eLife *3*, e03440.

Geertsma, E.R., Groeneveld, M., Slotboom, D.J., and Poolman, B. (2008). Quality control of overexpressed membrane proteins. Proc. Natl. Acad. Sci. USA *105*, 5722–5727.

Goder, V., and Spiess, M. (2003). Molecular mechanism of signal sequence orientation in the endoplasmic reticulum. EMBO J. *22*, 3645–3653.

Grisshammer, R., Duckworth, R., and Henderson, R. (1993). Expression of a rat neurotensin receptor in Escherichia coli. Biochem. J. *295*, 571–576.

Guglielmi, L., Denis, V., Vezzio-Vié, N., Bec, N., Dariavach, P., Larroque, C., and Martineau, P. (2011). Selection for intrabody solubility in mammalian cells using GFP fusions. Protein Eng. Des. Sel. *24*, 873–881.

Harley, C.A., Holt, J.A., Turner, R., and Tipper, D.J. (1998). Transmembrane protein insertion orientation in yeast depends on the charge difference across transmembrane segments, their total hydrophobicity, and its distribution. J. Biol. Chem. *273*, 24963–24971.

Heijne, G. (1986). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. *5*, 3021–3027.

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and von Heijne, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature *433*, 377–381.

Lewinson, O., Lee, A.T., and Rees, D.C. (2008). The funnel approach to the precrystallization production of membrane proteins. J. Mol. Biol. *377*, 62–73.

Noens, E.E., Williams, C., Anandhakrishnan, M., Poulsen, C., Ehebauer, M.T., and Wilmanns, M. (2011). Improved mycobacterial protein production using a Mycobacterium smegmatis groEL1ΔC expression strain. BMC Biotechnol. *11*, 27.

Ramasamy, S., Abrol, R., Suloway, C.J., and Clemons, W.M., Jr. (2013). The glove-like structure of the conserved membrane protein TatC provides insight into signal sequence recognition in twin-arginine translocation. Structure *21*, 777–788.

Rapoport, T.A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. Nature *450*, 663–669.

Rollauer, S.E., Tarry, M.J., Graham, J.E., Jääskeläinen, M., Jäger, F., Johnson, S., Krehenbrink, M., Liu, S.M., Lukey, M.J., Marcoux, J., et al. (2012). Structure of the TatC core of the twin-arginine protein transport system. Nature *492*, 210–214.

Sarkar, C.A., Dodevski, I., Kenig, M., Dudli, S., Mohr, A., Hermans, E., and Plückthun, A. (2008). Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc. Natl. Acad. Sci. USA *105*, 14808–14813.

Schlegel, S., Klepsch, M., Gialama, D., Wickström, D., Slotboom, D.J., and de Gier, J.W. (2010). Revolutionizing membrane protein overexpression in bacteria. Microb. Biotechnol. *3*, 403–411.

Scott, D.J., Kummer, L., Tremmel, D., and Plückthun, A. (2013). Stabilizing membrane proteins through protein engineering. Curr. Opin. Chem. Biol. *17*, 427–435.

Seppälä, S., Slusky, J.S., Lloris-Garcerá, P., Rapp, M., and von Heijne, G. (2010). Control of membrane protein topology by a single C-terminal residue. Science *328*, 1698–1700.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Tsirigos, K.D., Peters, C., Shu, N., Käll, L., and Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. *43*, W401–407.

Van Lehn, R.C., Zhang, B., and Miller, T.F., 3rd. (2015). Regulation of multispanning membrane protein topology via post-translational annealing. eLife *4*, e08697.

Wagner, S., Bader, M.L., Drew, D., and de Gier, J.W. (2006). Rationalizing membrane protein overexpression. Trends Biotechnol. *24*, 364–371.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. *17*, 691–695.

Wang, Z., Xiang, Q., Wang, G., Wang, H., and Zhang, Y. (2011). Optimizing expression and purification of an ATP-binding gene gsiA from Escherichia coli k-12 by using GFP fusion. Genet. Mol. Biol. *34*, 661–668.

Warne, T., Serrano-Vega, M.J., Baker, J.G., Moukhametzianov, R., Edwards, P.C., Henderson, R., Leslie, A.G.W., Tate, C.G., and Schertler, G.F.X. (2008). Structure of a beta1-adrenergic G-protein-coupled receptor. Nature *454*, 486–491.

Woodall, N.B., Yin, Y., and Bowie, J.U. (2015). Dual-topology insertion of a dual-topology membrane protein. Nat. Commun. *6*, 8099.

Zhang, B., and Miller, T.F., 3rd. (2010). Hydrophobically stabilized open state for the lateral gate of the Sec translocon. Proc. Natl. Acad. Sci. USA *107*, 5399–5404.

Zhang, B., and Miller, T.F., 3rd. (2012a). Direct simulation of early-stage Sec-facilitated protein translocation. J. Am. Chem. Soc. *134*, 13700–13707.

Zhang, B., and Miller, T.F., 3rd. (2012b). Long-timescale dynamics and regulation of Sec-facilitated protein translocation. Cell Rep. *2*, 927–937.