Cell Reports, Volume *16*

Supplemental Information

A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency

Stephen S. Marshall, Michiel J.M. Niesen, Axel Müller, Katrin Tiemann, Shyam M. Saladi, Rachel P. Galimidi, Bin Zhang, William M. Clemons, Jr., and Thomas F. Miller, III

Supplemental Experimental Procedures

Designing and cloning of TatC chimeras.

The parent plasmid used for cloning, pET28(a+)-GFP-ccdB, was derived from an IMP-GFP vector used by (Drew et al., 2001). TatC homologs and chimeras were prepared from genomic DNA, with the exception of wild-type *M. tuberculosis* and *A. aeolicus* TatC genes which were synthesized by primer extension as applied in DNAWorks (NIH) (Hoover and Lubkowski, 2002). In most cases, the Gibson assembly cloning protocol was used for cloning (Gibson et al., 2009). Expression of a vector containing *Aa*TatC with an N-terminal ten-His tag and without the GFP fusion-tag was used as a negative control for in-gel fluorescence, western blot analysis and flow cytometry. For constructs containing the β-lactamase tag, the GFP sequence was removed and replaced with a β-lactamase sequence using Gibson cloning. For generation of *M. smegmatis* compatible plasmids, the entire coding region of the TatC homologs including the entire GFP sequence and the poly-His tag were PCR amplified out of their respective pET28(a+)-GFP-ccdB vector using primers with compatible regions for placement into the pMyNT vector using Gibson assembly (Noens et al., 2011). For β-lactamase constructs, the GFP sequence was replaced by a β-lactamase sequence using Gibson assembly.

E. coli expression.

Plasmids were transformed into BL21 Gold (DE3) cells and transferred onto LB agar plates containing 50 μg/ml kanamycin plates after one-hour incubation. After overnight incubation at 37°C, colonies were scraped off the plates into 5 mL of LB, resuspended, and the OD_{600} was determined. These samples were then diluted into 50 mL 2xYT containing 50 μg/ml kanamycin in 125 mL baffled flasks to a starting OD_{600} of approximately 0.01. Cultures were grown in an orbital shaker at 37°C until they reached an OD_{600} of 0.15. The temperature of the orbital shaker was then reduced to 16°C. Upon reaching an OD_{600} of 0.3, IPTG was added to final

concentration of 1mM to induce expression. Cultures were grown for a further 16 hours prior to analysis.

β-lactamase survival test.

Plasmids containing the β-lactamase tag were expressed overnight at 16°C as previously described. Cells from each overnight culture were washed with phosphate buffered saline (PBS) to remove IPTG then diluted into fresh 50 mL 2xYT media containing 50 μg/ml kanamycin to a starting OD_{600} of 0.1 in 125 mL baffled flasks. Cultures were grown at 37°C to an OD_{600} of approximately 0.5 where a control sample from each culture was taken, diluted 10,000 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/ml kanamycin. To each culture, 50 μg/ml ampicillin was added and shaken at 37°C for a further 90 minutes. A sample from each culture was taken, diluted 200 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/ml kanamycin. Plates were grown overnight (~16 hours) and the number of colonies on each plate was counted. Colony counts from the second plating were normalized by the colony counts from the first plating to account for variation in the OD_{600} at which ampicillin was added to determine relative survival. The procedure was performed in triplicate and standard errors of normalized values were calculated. For each plot of relative survival, the values are normalized to the highest survival rate of the samples in the figure.

M. smegmatis expression.

For *M. smegmatis* overexpression, constructs were transformed into mc^2155 cells using electroporation and transferred onto Middlebrook 7H11 plates (10.25 g Middlebrook 7H11 Agar Base, 1 vial ADC growth supplement, 2.5 g glycerol, 1 mM $CaCl_2$, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, 50 μg/mL hygromycin, and water to 500 mL) after a three hour incubation in 1 mL Middlebrook 7H9 culture media (2.35 g Middlebrook 7H9 Broth Base, 1 vial ADC growth supplement, 0.5 g Tween-80, 1 mM $CaCl_2$, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, and water to 500 mL). Plates were grown for three to four days until colonies formed. Single colonies were picked into 5 mL Middlebrook 7H9 culture media containing 50

μg/mL hygromycin. The following day, 50 mL cultures of Middlebrook 7H9 expression media (2.35 g Middlebrook 7H9 Broth Base, 0.25 g Tween-80, 1 g glycerol, 1 g glucose, 1 mM $CaCl_2$, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, 50 μg/mL hygromycin, and water to 500 mL) were inoculated at a starting OD_{600} of 0.005. Cultures were grown at 37°C and expression was induced with 0.2% acetamide at an OD_{600} of 0.5. Cultures were grown for six hours after induction prior to analysis.

Flow cytometry.

A 200 μL sample of each expression culture was centrifuged at 4000g for 3 minutes to pellet the cells and then the supernatant was removed. Cells were resuspended in 1 mL of PBS and 200 μL of each were dispensed into 96-well plates and kept on ice for analysis. Whole-cell GFP fluorescence was determined using a MACSQuant10 Analyzer. Forward scattering, side scattering and total fluorescence at 488 nm were considered during analysis. Measured events were gated based on the negative control sample to contain the lowest 90% of both forward and side scattering values to remove anomalous particles, such as dead or clumped cells. Mean cell fluorescence was calculated for the gated population as a measure of folded TatC. At least four independent expression trials were performed for each sequence tested to ascertain expression variance. Flow cytometry data analysis was performed with FlowJo Software. Flow cytometry data is normalized to a standard for each day data was collected. For example, for '*Aa*-tail/wild-type' data points, the mean fluorescence of the *Aa*-tail swap chimeras were normalized by the mean fluorescence of their respective homologs containing the wild-type tail for that day's trial. Similarly, for relative fluorescence data points in which wild-type *Aa*TatC was the standard, the mean fluorescence of each sample was normalized by the mean fluorescence of the *Aa*TatC sample for that day's trial. In both cases, final calculated values are averages of the normalized values over at least four trials with error bars representing standard errors of the mean for those normalized values.

In-gel fluorescence and western blot analyses.

In-gel fluorescence and western blot analyses were used as an alternative measure of total expressed proteins. 5 mL of expression samples were centrifuged and supernatant discarded. Samples were resuspended to an OD_{600} of 3.0 in PBS. 1 mL of each sample was collected and 250 µL lysis buffer (375 mM Tris-HCl pH 6.8, 6% SDS, 48% glycerol, 9% 2-Mercaptoethanol, 0.03% bromophenol blue) was added. Samples were lysed via freeze fracturing by three rounds of freezing using liquid nitrogen and thawing using room temperature water. 20 µL of each lysed sample was subjected to SDS-PAGE. SDS-PAGE gels were imaged for fluorescence using a UV gel imager with a filter for GFP fluorescence to determine in-gel fluorescence.

For western blot analysis, the samples were transferred from the gel onto a nitrocellulose membrane using the Trans-Blot Turbo System. The membranes were washed three times with 15 mL TTBS (50 mM Tris pH 7.6, 150 mM NaCl, 0.05% Tween-20), incubated one hour with 15 mL 5% milk powder in TTBS, washed three times with 15 mL of TTBS, then incubated with 1:5000 anti-GFP Mouse primary antibody (EMD Millipore, Lot # 2483215) in 15 mL 5% milk powder in TTBS overnight. Membranes were washed three times with 15 mL TTBS, incubated with 1:15000 IRDye® 800CW Donkey anti-Mouse secondary antibody (LI-COR, Lot # C31024-04) in 15 mL 5% milk powder in TTBS for one hour, washed three times with 15 mL TTBS, then visualized using a Licor IR western blot scanner. ImageJ was used to process the images.

The CG simulation model: Overview.

The CG model is employed with only minor modifications from (Zhang and Miller, 2012), all of which are specified below. Key features of the CG model and its implementation are provided here; for a full discussion of the CG model, the reader is referred to (Zhang and Miller, 2012).

As described in (Zhang and Miller, 2012), the CG model explicitly describes the configurational dynamics of the nascent-protein chain, conformational gating in the Sec translocon, and the slow dynamics of ribosomal translation. The nascent chain is represented as a

freely jointed chain of beads, where each bead represents three amino acids and has a diameter of 8 Å, the typical Kuhn length for polypeptide chains (Hanke et al., 2010, Staple et al., 2008). Bonding interactions between neighboring beads are described using the finite extension nonlinear elastic (FENE) potential (Kremer and Grest, 1990), short-ranged nonbonding interactions are modeled using the Lennard-Jones potential, electrostatic interactions are modeled using the Debye-Hückel potential, periplasmic binding is included as described in (Zhang and Miller, 2012) for BiP, and solvent interactions are described using a position-dependent potential based on the water-membrane transfer free energy for each CG bead; all parameters are the same as used previously (Zhang and Miller, 2012), unless otherwise stated. The time evolution of the nascent protein is modeled using overdamped Langevin dynamics, with the CG beads confined to a two-dimensional subspace that runs along the axis of the translocon channel and between the two helices of the lateral gate (LG). Conformational gating of the translocon LG corresponds to the LG helices moving out of the plane of confinement for the CG beads, allowing the nascent chain to pass into the membrane bilayer. The rate of stochastic LG opening and closing is dependent on the sequence of the nascent protein CG beads that occupy the translocon channel. Ribosomal translation is directly simulated via growth of the nascent protein at the ribosome exit channel; throughout translation, the C-terminus of the nascent protein is held fixed, and new beads are sequentially added at a rate of 24 residues per second. Upon completion of translation, the C-terminus is released from the ribosome. It has been confirmed that the results presented in the current study are robust with respect to changes in the rate of ribosomal translation (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained using a rate of translation of 24 residues/sec and 6 residues/sec, $r = 0.99 \pm 0.06$).

The CG simulation model: Implementation details.

Two changes to the protocol for the CG simulation model were introduced in the current study, with respect to the protocol used in (Zhang and Miller, 2012). These modifications were included to remove unphysical artifacts in the simulations, although it is emphasized that

conclusions in the main text are qualitatively unchanged by these modifications (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained with and without the modifications to the simulation protocol, $r = 0.97 \pm 0.09$).

The first change in the CG model is that the ribosome is assumed to remain associated with the translocon following translation of the nascent protein. In the previously implementation of the model, the ribosome was assumed to dissociate from the translocon immediately following stop-translation, which was found in the current study to lead to artifacts for nascent proteins with extremely short C-terminal domains. Furthermore, this modification is consistent with experimental evidence that indicates that the timescale for ribosomal dissociation is slower than the trajectories simulated here (Schaletzky and Rapoport, 2006, Potter and Nicchitta, 2002).

The second change in the CG model relates to the potential energy cost of flipping hydrophilic nascent-protein loops across the lipid membrane at significant distances from the translocon. The Wimley-White water-octanol transfer free energy scale (Wimley et al., 1996) that was used to parameterize the interactions of the CG beads with the membrane is appropriate for describing the transfer of amino acids between an aqueous region and either the phospholipid interface or the region of the membrane interior that is close to the translocon lateral gate (MacCallum and Tieleman, 2011). However, the flipping of hydrophilic nascent-protein loops across the membrane at significant distances from the translocon involves moving CG beads through the hydrophilic core of the membrane interior, which will incur a large potential energy barrier (MacCallum and Tieleman, 2011). To account for this effect, and to avoid unphysical flipping of short hydrophilic loops across the membrane, an additional potential energy term was included in potential energy function that describes the interactions between the CG beads and the membrane,

$$U_{core}(x, y) = g S(x; \phi_x, \psi_x)\left[1 - S(y; \phi_y, \psi_y)\right], \quad [1]$$

$$S(x; \phi_x, \psi_x) = \frac{1}{4}\left(1 + \tanh\frac{x - \phi_x}{b}\right)\left(1 + \tanh\frac{x - \psi_x}{b}\right) \quad [2]$$

where $\phi_x = -1\sigma$, $\psi_x = 1\sigma$, $\phi_y = -2.5\sigma$, $\psi_y = 2.5\sigma$, and $b = 0.25\sigma$. The parameters σ and g are respectively the diameter of the CG beads and the water-octanol transfer free energy for the CG beads, both of which appear in the original model. We emphasize that this new term has no noticeable effect on the potential energy function for the CG beads at distances within 8 Å to the translocon channel; it simply affects unphysical flipping of the TM domains across the membrane at larger distances from the channel. This artifact was not observed in the earlier study using the CG model, since only processes involving the translocation or membrane integration of a single TM domain were considered.

The CG simulation model: Mapping IMP amino-acid sequence to the CG model.

In the current study, amino-acid sequences for the TatC homologs are mapped onto sequences of CG beads as follows. Each consecutive trio of amino acid residues in the nascent protein sequence is mapped to an associated CG bead. The water-membrane transfer free energy for each CG bead is taken to be the sum of the contributions from the individual amino acids; these values are taken from the experimental water-octanol transfer free energies for single residues (Wimley et al., 1996). The charge for each CG bead is taken to be the sum of the contribution from the individual amino acids. As in (Zhang and Miller, 2012), positively charged residues (arginine and lysine) were modeled with a +2 charge to capture significant effects on topology due to changes in the nascent protein sequence. Histidine residues were modeled with a +1 charge to account for the partial protonation of these residues, and negatively charged residues (glutamate and aspartate) were modeled with a charge of -1. The mapping procedure for *Aa*TatC is depicted in Figure 3A as an example.

In the MtTatC chimeras where loop 5 was replaced (Figure 7), the mapping protocol was modified to avoid a frame-shift in the three-to-one mapping of amino acids. Specifically, prior to mapping amino acids to beads as described previously, 0-2 dummy amino acids were added to the sequence immediately following loop 5. The number of dummy amino acids was chosen such that the amino acid to bead mapping was identical to that of *Mt*TatC wildtype for TMD 6

onwards, avoiding a frame-shift. Dummy amino acids have zero charge and zero water-membrane transfer free energy.

The CG simulation model: Calculation details.

For the results in Figures 3-6, the co-translational membrane integration for each TatC sequence is simulated using 1200 independent CG trajectories; for the results in Figure 7, each sequence is simulated using over 400 independent trajectories. As in (Zhang and Miller, 2012), each CG trajectory is performed with a timestep of 100 ns. All trajectories were terminated 30 seconds after the end of translation for the protein sequence.

The CG simulation model: Analysis of simulation results.

To determine whether a given trajectory leads to integration in the correct multispanning topology, the topology of a nascent protein configuration can be characterized by the location of the soluble loops that connect the TMD. We thus specify a collective variable λ_i associated with each loop, with i=1 corresponding to the loop that leads TMD 1 in the sequence (*i.e.* the N-terminal sequence) and i=7 corresponding to the loop that follows TMD 6 (*i.e.* the C-tail). If loop i is in the cytosol, then $\lambda_i = 1$; if loop i is in the periplasm, then $\lambda_i = -1$; otherwise, $\lambda_i = 0$. Whether a given loop is in the cytosol, in the membrane, or in the periplasm is determined by the tracking position of a representative bead in that loop (Table S2). Representative beads were chosen based on having the lowest probability of being inside the lipid region compared to other beads in that loop. A given trajectory is determined to have reached correct IMP integration ($\lambda_i = -1$ for periplasmic loops and, $\lambda_i = 1$ for cytosolic loops) if a configuration with the loops in the correct orientation is sampled during a time window of 6 seconds taken 25 seconds after the end of translation; the time window of 25 seconds was found sufficient to allow the nascent protein to finish the integration/translocation of TMD 6.

Figure S4 shows the fraction of trajectories that exhibit the correct topology for each individual loop for all TatC homologs and chimeras considered in this study. It is clear from

Figure S4 that the changes to the amino-acid sequence considered in this study largely only impact the topology of the domain where the changes to the amino acid sequence were introduced; the topology of the rest of the protein is not predicted by the CG simulation model to be significantly affected by the sequence changes. The calculated results are robust with respect to the details of the definition of simulated integration efficiency (Pearson correlation coefficient between Wt/Mutant ratios obtained analyzing only the loop that was modified and those obtained analyzing all loops, $r = 0.85 \pm 0.16$) (Figure S3); to minimize statistical error, for all simulation results presented in the main text, the topology of the IMP is thus characterized in terms of only the loop of interest. Specifically, results in Figure 3-6 report on loop 7 and results in Figure 7 report on loop 5.

References

DREW, D. E., VON HEIJNE, G., NORDLUND, P. & DE GIER, J. W. 2001. Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli. *FEBS Lett,* 507**,** 220-4.

GIBSON, D. G., YOUNG, L., CHUANG, R. Y., VENTER, J. C., HUTCHISON, C. A., 3RD & SMITH, H. O. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods,* 6**,** 343-5.

HANKE, F., SERR, A., KREUZER, H. J. & NETZ, R. R. 2010. Stretching single polypeptides: The effect of rotational constraints in the backbone. *EPL,* 92.

HOOVER, D. M. & LUBKOWSKI, J. 2002. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.,* 30**,** e43.

KREMER, K. & GREST, G. S. 1990. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J Chem Phys,* 92**,** 5057-5086.

MACCALLUM, J. L. & TIELEMAN, D. P. 2011. Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions. *Trends Biochem Sci,* 36**,** 653-62.

NOENS, E. E., WILLIAMS, C., ANANDHAKRISHNAN, M., POULSEN, C., EHEBAUER, M. T. & WILMANNS, M. 2011. Improved mycobacterial protein production using a Mycobacterium smegmatis groEL1DeltaC expression strain. *BMC Biotechnol,* 11**,** 27.

POTTER, M. D. & NICCHITTA, C. V. 2002. Endoplasmic reticulum-bound ribosomes reside in stable association with the translocon following termination of protein synthesis. *J Biol Chem,* 277**,** 23314-20.

SCHALETZKY, J. & RAPOPORT, T. A. 2006. Ribosome binding to and dissociation from translocation sites of the endoplasmic reticulum membrane. *Mol Biol Cell,* 17**,** 3860-9.

STAPLE, D. B., PAYNE, S. H., REDDIN, A. L. & KREUZER, H. J. 2008. Model for stretching and unfolding the giant multidomain muscle protein using single-molecule force spectroscopy. *Phys Rev Lett,* 101**,** 248301.

WALDO, G. S., STANDISH, B. M., BERENDZEN, J. & TERWILLIGER, T. C. 1999. Rapid protein-folding assay using green fluorescent protein. *Nat Biotechnol,* 17**,** 691-5.

WIMLEY, W. C., CREAMER, T. P. & WHITE, S. H. 1996. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry,* 35**,** 5109-24.

ZHANG, B. & MILLER, T. F., 3RD 2012. Long-timescale dynamics and regulation of Sec-facilitated protein translocation. *Cell Rep,* 2**,** 927-37.

SI Figures and Tables

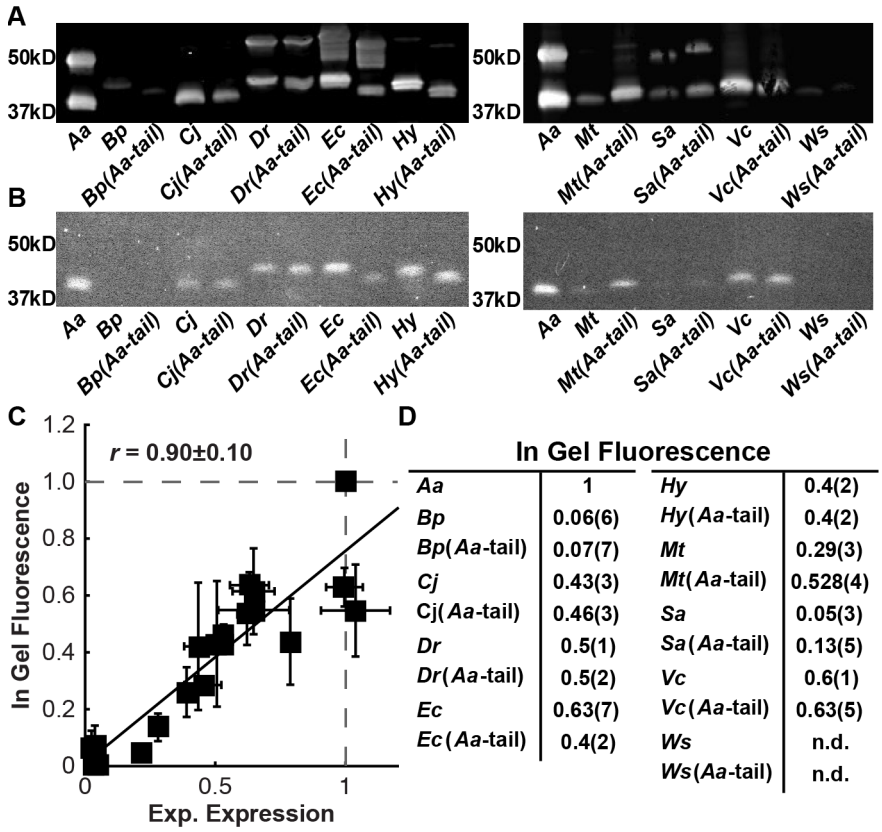

Figure S1. Related to Figures 1 and 2. Additional experimental data for expression of TatC variants in *E. coli*. (**A**) Anti-GFP western blot results for TatC homologs and the corresponding *Aa*-tail swap chimeras. Two bands were observed for all lanes where TatC-GFP was at high relative concentrations with the lower bands active by in-gel fluorescence and therefore determined to be folded protein. (Waldo et al., 1999) (**B**) In-gel fluorescence of SDS-PAGE for TatC homologs and the corresponding *Aa*-tail swap chimeras. Bands that exhibit fluorescence represent folded protein. The results exhibit the same trends in expression yield as seen by flow-cytometry. (**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 (Pearson correlation coefficient r=0.9±0.1) with in-gel fluorescence showing the same trends in expression yield as seen by flow-cytometry. Error bars indicate the standard error

of mean. (**D**) Average in-gel fluorescence quantified across four separate gels. *Ws* and *Ws*(*Aa*-tail) could not be detected (n.d.) by in-gel fluorescence. Values for each band are normalized to the *Aa*TatC band and values in parentheses indicate the standard error of mean.

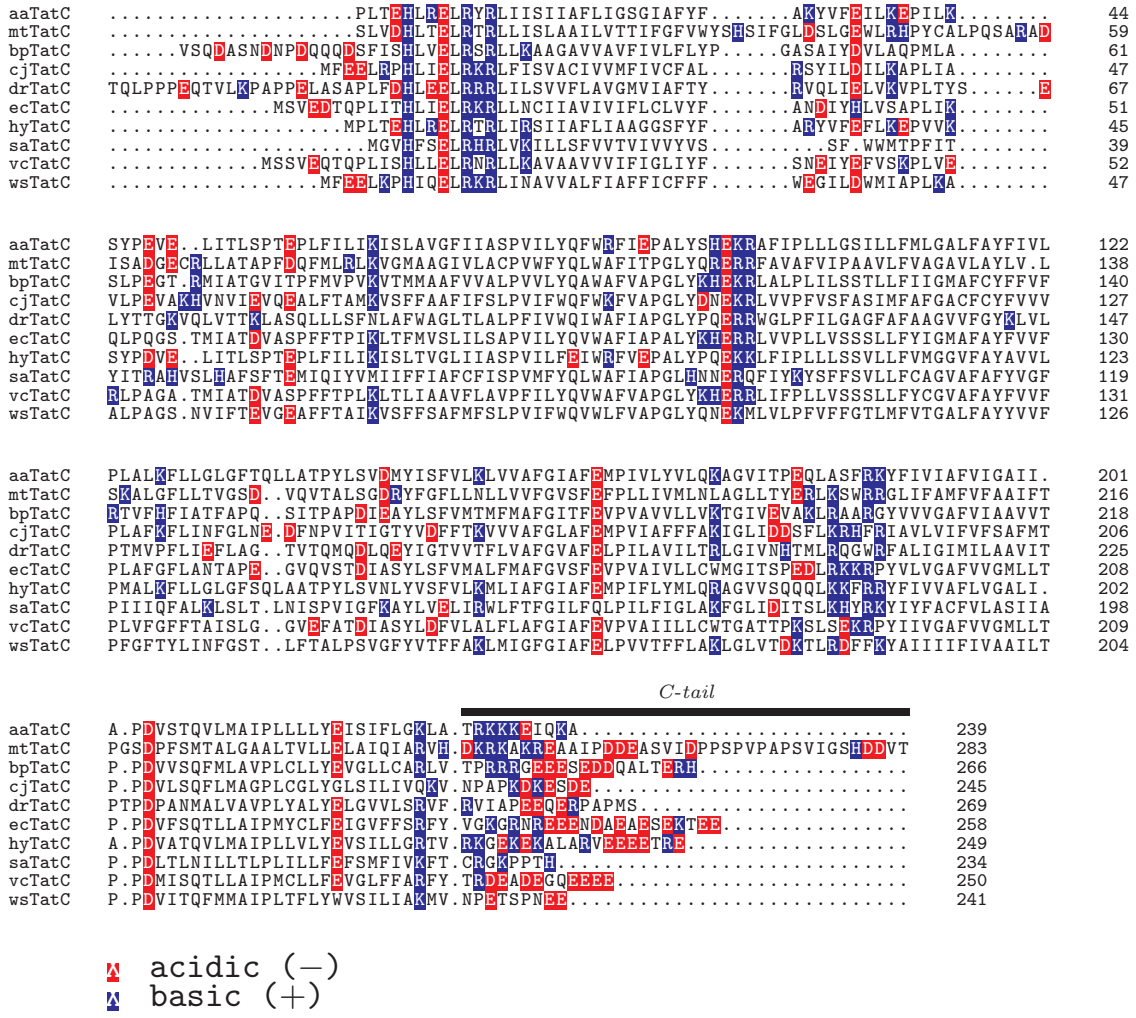

Figure S2. Related to Figure 1. Sequence alignment. Sequence alignment of wild-type TatC homologs. Clustal Omega was used to align sequences.

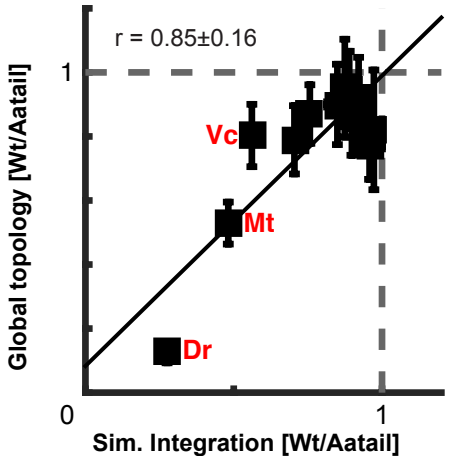

Figure S3. Related to Figure 3. Robustness of the simulation results with respect to the definition of simulated integration efficiency. Correlation of the simulated integration efficiency calculated using only the loop that was modified (x-axis) versus the simulated integration efficiency calculated using the full multispanning topology (y-axis). Both metrics are highly correlated (Pearson correlation coefficient r=0.85), but use of only the modified loop avoids statistical error due to fluctuations in the topology of the remaining loops. This figure includes data of all the chimeras that were computationally studied, including those presented in Figures 4-9. Error bars indicate the standard error of mean.

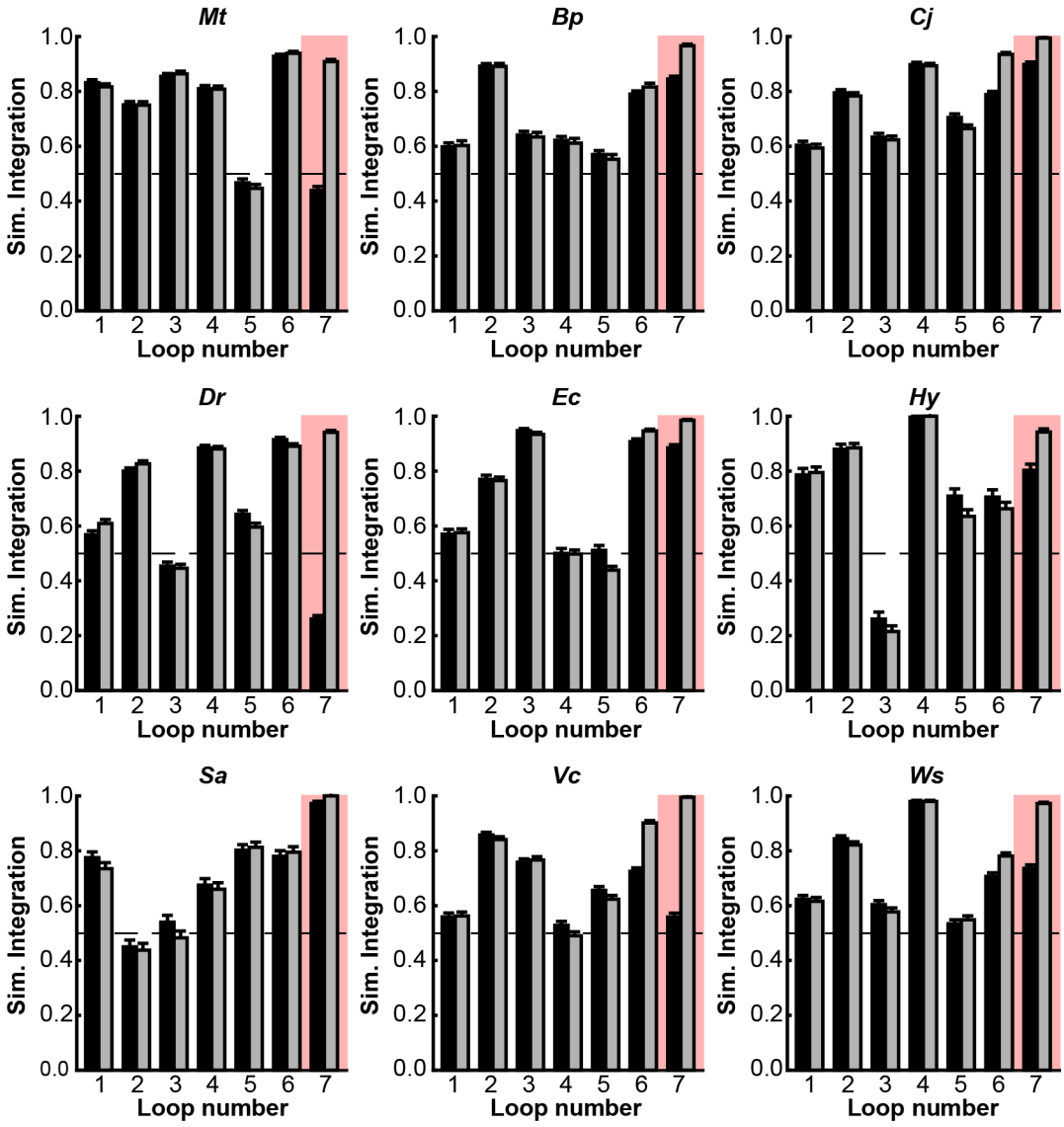

Figure S4. Related to Figure 3. Simulated integration efficiency calculated for each loop in the tested TatC wildtypes and *Aa*-tail chimeras. For each considered TatC homolog, the simulated integration efficiency for the individual loops for both the wild-type sequence (black bars) and the *Aa*-tail chimeras (grey bars). It is seen that the *Aa*-tail generally leads to a significant effect on the integration efficiency of loop 7 (highlighted), with smaller effects on the other loops. Error bars indicate the standard error of mean.

Table S1. Related to Figures 1-7. DNA and protein sequences of all wild-type and mutant TatCs tested, excluding C-terminal tags.

	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop 6	Loop 7
AaTatC	7-9	43-45	88-90	145-147	181-183	202-204	238-239
Mt	7-9	61-63	112-114	151-153	193-195	220-222	244-246
Mt(Aa-tail)	7-9	61-63	112-114	151-153	193-195	220-222	244-246
Bp	25-27	64-66	112-114	160-162	196-198	220-222	253-255
Bp(Aa-tail)	25-27	64-66	112-114	160-162	196-198	220-222	250-252
Cj	13-15	55-57	100-102	139-141	187-189	208-210	238-240
Cj(Aa-tail)	13-15	55-57	100-102	139-141	187-189	208-210	238-240
Dr	28-30	73-75	118-120	166-168	202-204	229-231	262-264
Dr(Aa-tail)	28-30	73-75	118-120	166-168	202-204	229-231	247-249
Ec	10-12	55-57	103-105	142-144	190-192	211-213	244-246
Ec(Aa-tail)	10-12	55-57	103-105	142-144	190-192	211-213	244-246
Hy	7-9	40-42	94-96	139-141	184-186	205-207	232-234
Hy(Aa-tail)	7-9	40-42	94-96	139-141	184-186	205-207	232-234
Sa	7-9	43-45	91-93	142-144	178-180	199-201	229-231
Sa(Aa-tail)	7-9	43-45	91-93	142-144	178-180	199-201	229-231
Vc	16-18	52-54	103-105	145-147	190-192	211-213	247-249
Vc(Aa-tail)	16-18	52-54	103-105	145-147	190-192	211-213	241-243
Ws	10-12	61-63	97-99	148-150	181-183	205-207	241
Ws(Aa-tail)	10-12	61-63	97-99	148-150	181-183	205-207	235-237

Table S2. Related to Figure 3. Loop definitions used in simulation trajectory analysis. Each loop is specified in terms of the amino-acid residue sequence numbers (end-points inclusive) associated with the wild-type sequence. Complete description of the loop topology analysis is provided in the Supplemental Experimental Procedures.