PNAS
www.pnas.org

Supplementary Information for

An enzyme-based biosensor for monitoring and engineering protein stability in vivo

Chang Ren, Xin Wen, Jun Mencius, and Shu Quan*

Shu Quan
Email: shuquan@ecust.edu.cn

This PDF file includes:

Supplementary: Material and Methods
Figures S1 to S13
Tables S1 to S3
SI References

Supplementary Information Text

Materials and Methods

Plasmids. All plasmids used in this work are listed in **Table S3**. *E. coli* strain Trans1T1 (TransGen Biotech) was used for cloning. Plasmids construction in this work follows standard restriction enzyme cloning technique or overlap extension PCR cloning technique as described previously (1). The portion of *cysG* encoding urogen III methyltransferase ($CysG^A$) (2) was amplified from the genome of *E. coli* MG1655 and subsequently cloned into pTrc99a derived vector pssTrx. For the identification of permissive sites in $CysG^A$, we constructed the plasmids pTrc-$CysG^A$-insertion 1-10 containing the Im7 protein flanked by two glycine-serine (GS) linkers (SSGSSG and GGGGSGGGGS) at 10 candidate permissive sites, respectively. The Im7-encoding gene was amplified from pCDFTrc-ssIm7 WT (3). For expression of the $CysG^A$ tripartite fusions containing the MBP, the MBP-encoding gene (*malE*) without the signal sequence was amplified from genomic DNA of *E. coli* MG1655 and replaced the Im7-encoding gene in the plasmid pTrc-$CysG^A$-Im7 (Im7 inserted after the G364 site of $CysG^A$), resulting in the plasmid pTrc-$CysG^A$-MBP. For expression of the $CysG^A$ tripartite fusions containing the AcP, polyQ tracts, hIAPP, and $MLL3_{SET}$, the same approach as described above was followed. The AcP-encoding gene and the $MLL3_{SET}$-encoding gene were amplified from pDONR233-AcP and pET28b-His-SUMO-$MLL3_{SET}$, respectively (both kindly provided by Y. Chen, Chinese Academy of Sciences). Different lengths of polyQ-encoding sequence were amplified from pBAD33-Q20, pBAD33-Q45 and pBAD33-Q87 (all kindly provided by J. Bardwell, University of Michigan). The hIAPP-encoding sequence was synthesized by GENEWIZ. For expression and purification of $CysG^A$, AcP variants, and $MLL3_{SET}$ variants, the corresponding DNA sequences were cloned into pET28a vector containing an N-terminal His-SUMO tag with a ULP1 cleavage site. Point mutations of the genes were generated using standard site-directed mutagenesis. All constructs were confirmed by sequencing.

$CysG^A$ fluorescence assay and fractionation of *E. coli* extract. Plasmids containing the $CysG^A$ tripartite system were transformed into *E. coli* strain Trans1T1. A single colony from the fresh cells was inoculated into 4-mL LB medium (10 g/L tryptone, 5 g/L yeast

extract, 10 g/L NaCl) supplemented with 200 μg/mL ampicillin. Cells were grown overnight at 37 °C with shaking. Overnight cultures were diluted 1:100 in LB medium supplemented with 200 μg/mL ampicillin and grown at 37 °C with shaking until OD_{600} reached 0.5. Expression of fusion proteins was induced with isopropyl-β-D-thiogalactoside (IPTG) at a final concentration of 10 μM. The culture was then incubated at 37 °C for 5 h. For the measurement of fluorescence, cells were collected and resuspended in phosphate buffered saline (PBS) to normalize OD_{600} to 1.0. Then, 100 μL of the culture was transferred to the wells of 96-well plates. The fluorescence of each well was measured using an automated plate reader (SynergyHTX hybrid reader, Bio-Tek) by fixing the excitation filter at 380 ± 20 nm and the emission filter at 600 ± 40 nm.

For soluble protein extraction, cells were collected and resuspended in buffer A (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 0.5 mg/mL lysozyme, 0.05 mg/mL DNaseI) to obtain an OD_{600} of 1.0. The cells were lysed by freezing and thawing as described previously (4). After centrifugation for at 16,000 g for 20 min, the supernatant was considered as the soluble fraction. For fluorescence spectra analysis, 1 mL of supernatant was added into a quartz cuvette and excited at 357 nm while recording emission spectra from 500-700 nm or excited at 300-450 nm while recording excitation spectra on emission at 620 nm using a Lumina Fluorescence Spectrometer (Thermo Scientific).

Protein expression and purification. A single colony of *E. coli* BL21 (DE3) harboring pET28b-His-SUMO-$CysG^A$ was picked and inoculated into 10 mL of LB medium. The overnight culture was diluted into 1 L of LB medium supplemented with 100 μg/mL kanamycin and grown at 37 °C until OD_{600} reached 0.6. The culture was shifted to 16 °C and induced with 0.1 mM IPTG for 16 h. Cells were harvested by centrifugation at 4 °C for 30 min and resuspended in 40 mL of lysis buffer A (50 mM NaP, pH 8.0, 400 mM NaCl, 10% glycerol, 0.1 mg/mL lysozyme, and protein inhibitor cocktail). The suspension was subjected to a high-pressure cell disruptor for 10 min at 800 psi. Cell debris was removed by centrifugation at 10,000 g for 60 min, and then the supernatant was incubated with 2-mL cOmplete His-Tag Purification Resin (Roche) at 4 °C for 2 h with gentle rotation. The resin was collected and washed with lysis buffer followed by elution with 300 mM imidazole in lysis buffer. The SUMO tag was removed by

incubation with ULP1 protease at 4 °C for 2 h. The resulting protein was concentrated and further purified by HiLoad 16/600 Superdex 75pg column (GE Healthcare) in gel filtration buffer (25 mM Tris-HCl, pH 8.0, 300 mM NaCl, 10% glycerol). Purified $CysG^A$ was used to develop polyclonal anti- $CysG^A$ antibody (prepared by HangZhou HuaAn Biotechnology Co., Ltd).

$MLL3_{SET}$ and its variants were expressed and purified as described previously with minor modifications (5). *E. coli* BL21 (DE3) harboring pET28b-His-SUMO-$MLL3_{SET}$ with different variant sequences was cultured as described above except that 10 μM $ZnSO_4$ was also added to each liter of LB medium to provide the ligand of $MLL3_{SET}$. The cells were induced with 0.1 mM IPTG at 16 ℃ for 14 h. Cells were collected by centrifugation at 4 ℃ for 30 min and then resuspended in 40 mL of lysis buffer B (50 mM Tris-HCl, pH 8.0, 400 mM NaCl, 10% glycerol, 2 mM β-mercaptoethanol, and protease inhibitor cocktail). The suspension was subjected to the high pressure cell disruptor for 10 min at 800 psi. Cell debris was removed by centrifugation at 10,000 g for 60 min, and then supernatant was incubated with 2-mL cOmplete His-Tag Purification Resin (Roche) at 4 °C for 2 h with gentle rotation. The protein was then eluted by on-beads digestion with ULP1 protease, followed by gel-filtration chromatography on HiLoad 16/600 Superdex 75pg column (GE Healthcare) in gel filtration buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl and 10% glycerol). Purified proteins of >95% purity as analyzed by SDS-PAGE were pooled, concentrated, and flash-frozen in liquid nitrogen before storing at −80 °C.

Western blot analysis. A single colony of *E. coli* Trans1T1 bearing different constructs was used to inoculate 4-mL LB medium supplemented with 200 μg/mL ampicillin at 37 ℃ overnight with shaking at 220 rpm. Overnight cultures were diluted 1:100 in LB medium and grown at 37 °C with shaking until OD_{600} reached 0.5. The expression of fusion proteins was induced with IPTG at a final concentration of 10 μM. The culture was then incubated at 37 °C for 5 h. The induction of $CysG^A$-$MLL3_{SET}$ and its variants was performed at 30 ℃ for 9 h. Then, cells were harvested by centrifugation at 4 ℃ and 12,000 rpm for 5 min and resuspended in phosphate buffered saline (PBS) to normalize OD_{600} to 2.0. The freeze-thaw method was used to lyse the cells. After centrifugation at

16,000 g for 20 min, the supernatants were collected as the soluble fraction. The soluble fractions were mixed with 5 × protein loading buffer and subjected to SDS-PAGE after boiled for 10 min.

For western blotting, proteins were transferred to polyvinylidene fluoride (PVDF) membranes (Merck) using semi-dry transfer apparatus (Tanon). Membranes were blocked with 5% non-fat milk in TBST (Tris buffer saline containing 0.1% Tween-20). Membranes were then incubated with rabbit anti-$CysG^A$ or rabbit anti-Trigger factor (1:2,000, GenScript, Cat. No. A01329) in TBST containing 5% non-fat milk, followed by incubation with fluorescently labeled IRDye 800 CW secondary antibodies (1:10,000, LI-COR Biosciences). Detection and quantitative analysis were performed by using Odyssey Sa (LI-COR Biosciences). For quantification, the ratio of target protein to anti-Trigger factor protein was calculated and normalized.

Construction and screening of random mutagenesis library. Error-prone PCR (Agilent Technologies, GeneMorph II Random Mutagenesis Kit) was used to construct the random mutagenesis libraries of AcP, IAPP, and $MLL3_{SET}$. The mutant fragments were inserted into $CysG^A$ by restriction digestion (with BamHI and XhoI from Thermo Scientific) and ligation (with T4 ligase from New England Biolabs). For the library of AcP and IAPP, 2.5 ng of the template DNA was used, and a mutation frequency of 1-2 amino acids per clone was obtained. For the library of $MLL3_{SET}$, 5 ng of the template DNA was used to achieve a mutation frequency of 3-4 amino acids. Mutated fragments were ligated back into pTrc-$CysG^A$ or pTrc-MBP- $CysG^A$ (for AcP fragments). After ethanol precipitation, the ligation products were electroporated into *E. coli* strain NEB10β (New England Biolabs) competent cells. The cells expressing $CysG^A$-POI fusion proteins were spread on LB agar plates supplemented with 200 μg/mL ampicillin and incubated at 37 ℃ for 12 h. Then, all colonies were scraped from the plate surface and used for plasmid extraction. 20 colonies from each library were sequenced to estimate mutation frequency. For $MLL3_{SET}$ library, four independent libraries were mixed at an equimolar ratio to form the P_0 library.

For the screening, the random libraries were transformed into *E. coli* strain Trans1T1 and spread on LB plates containing 10 μM IPTG and 200 μg/mL ampicillin. After

incubation at 37℃ for 12 h (AcP and IAPP libraries) or 30 ℃ for 36 h ($MLL3_{SET}$ library), the plates were quickly photographed under UV light, and colonies with higher fluorescence were picked by hand. After restreaking 1,374 colonies picked from the $MLL3_{SET}$ library, mixed plasmids were extracted to constitute the P_1 library.

High throughput sequencing and data processing. To sequence the P_0 and P_1 libraries, a 535 base-pair (bp) fragment containing the $MLL3_{SET}$ gene (462 bp) and flanks of $MLL3_{SET}$ (the GS linkers, 73 bp) was amplified by using Phanta Max Super-Fidelity DNA polymerase (Vazyme). The same region from the construct containing wild-type $MLL3_{SET}$ was also amplified to serve as the control sample which was used to probe the background error in high throughput sequencing. Paired-end DNA sequencing was used to read both directions of the target fragment (535 bp) on Illumina MiSeq. Each base was recognized by CASAVA 1.8.2 and converted into a readable FASTQ format. For data quality control, Cutadapt 1.9.1 (6) was then used to remove low-quality reads with more than 10% ambiguous bases or sequence length shorter than 75 bp. Finally, sequencing data from both directions were assembled by Pandaseq 2.7 (7) to form complete sequences. These raw sequences were then used as the input for data analysis.

In data processing, we extracted the exact sequence of $MLL3_{SET}$ (462 bp) in each raw sequence based on alignments to the first and last 10 bps of the wild type $MLL3_{SET}$ sequence. Frame-shift mutants were also discarded in this process. 379,808 sequences of $MLL3_{SET}$ from the P_0 library and 919,626 sequences of $MLL3_{SET}$ from the P_1 library were obtained. By dividing the number of occurrence of a specific amino acid by the total reads at that position in the P_0 or P_1 library, the frequency of each amino acid substitution at each position in the $MLL3_{SET}$ sequence was calculated. It is worth noting that we applied a stringent quality filter: if the same amino acid mutation occurs less than ten times at a given site, the mutation will be excluded from the analysis. Ultimately, we identified 1,784 out of 3,080 different possible substitutions in P_0 library. To describe the stability landscape of $MLL3_{SET}$, we introduced the terms stability score and enrichment score.

Stability score represents the site-specific mutational tolerance. Its definition is given in the following equation:

$$\text{Stability score} = log_2 \frac{f_{P1,A}}{f_{P0,A}}$$

where amino acid *A* is the wild-type residue at each site of $MLL3_{SET}$, $f_{P1,A}$ and $f_{P0,A}$ are the frequencies of amino acid *A* in the hit and full libraries, respectively.

The enrichment score is used to quantify the enrich degree of each amino acid substitution at each position before and after screening, which is given by the following equation:

$$\text{Enrichment score} = f_{P1,S} log_2 \frac{f_{P1,S}}{f_{P0,S}}$$

Where amino acid S is the substituting residue at each site of $MLL3_{SET}$. If either $f_{P1,S}$ or $f_{P0,S}$ is equal to zero, the specific amino acid enrichment score at that specific site is regarded as 0. The accumulated enrichment score of each site was introduced to evaluate their overall potency of being stabilized. The accumulated enrichment score is given by the following equation:

$$\text{Accumulated enrichment score} = \sum_S f_{P1,S} log_2 \frac{f_{P1,S}}{f_{P0,S}}$$

Equilibrium unfolding experiments. The thermodynamic stabilities of the variants were obtained by determining the equilibrium urea denaturation curves as described previously (8). Purified AcP variants were dialyzed in 50 mM acetate buffer (pH 5.5, 28 °C), and then the protein stock was diluted to 0.2 mg/mL into sample solutions containing different urea concentrations in 50 mM acetate buffer (pH 5.5, 28 °C). Purified $MLL3_{SET}$ variants was directly diluted into 25 mM Tris-HCl, 300 mM NaCl, 10 % glycerol, 0-8 M urea, pH 8.0 to the final concentration of 1 μM. After equilibrating at 28 °C for 3 h, intrinsic tryptophan fluorescence of the samples was measured upon excitation at 280 nm and emission at 335 nm using a Lumina Fluorescence Spectrometer (Thermo Scientific). For $MLL3_{SET}$ variants, the tryptophan fluorescence of different samples was measured at 20 °C. By fitting the urea unfolding curve according to Santoro and Bolen (9), the parameters of the thermodynamic stability of different protein variants can be obtained. These parameters include the free energy of folding in the absence of denaturants ($\Delta G°_{UN}$), the dependence of $\Delta G°_{UN}$ on denaturant concentration (m value) and the urea concentration at which half of the protein molecules are denatured (C_m). In this work, an

averaged value of m was used for the same type of protein to determine $\Delta G°_{UN}$ more accurately.

Thioflavin T assays. Human islet amyloid polypeptide (IAPP) and its variants were synthesized and purchased from Genscript and Synpeptide with >95% purity. Thioflavin T assays were performed as described previously (10). Briefly, all peptides were dissolved in 100% Hexafluoroisopropanol (HFIP) (Sigma-Aldrich, Cat. No. 105228-25G) to obtain a final concentration of 1 mM. The peptides were then diluted to a final concentration of 10 μM into 20 mM sodium acetate, pH 6.5 and 10 μM Thioflavin T. Fluorescence was monitored at 37 °C with agitation and was excited at 440 nm with emission at 485 nm using a Lumina fluorescence spectrometer (Thermo Scientific).

Relative surface accessibility (RSA). RSA values were calculated by dividing the absolute solvent accessible surface area for each residue by the maximum possible area for the amino acid type. Data of the absolute solvent accessible surface area for each residue was taken from the crystal structure of $MLL3_{SET}$ (PDB: 5F59). The maximum possible area for the amino acid type was calculated as described (11)

Site-saturation mutagenesis. To compensate for the coding bias caused by error-prone PCR, we selected 11 sites with top accumulated enrichment scores to construct their site-saturation mutagenesis libraries. The primers contain an NNK degenerate codon (N=A/T/G/C, K=G/T) at the sites subjected to mutagenesis. The screening procedures for these site-saturation mutagenesis libraries are the same as those described in the section "Construction and Screening of Random Mutagenesis Library".

Analysis of the SET domain sequences in the TRX/MLL family. Protein sequences of the TRX/MLL family were retrieved from UniProt. The SET domain sequences were extracted from the full-length protein sequences using a Python script. Sequence identity of each extracted SET domain sequence compared with $MLL3_{SET}$ was calculated using Clustal Omega (12). Sequences with identity less than 30% were discarded. The remaining 744 sequences were aligned using the MUSCLE 3.8.31_i86win32 multiple sequence alignment program (13). The frequency of each amino acid at each $MLL3_{SET}$ site was calculated using another python script, while gaps in the alignment were

excluded from the analysis. The sequence conservation at a given site in the multiple sequence alignment (MSA) is given by the following equation:

$$\text{Sequence conservation score} = \sum_{a} f_{msa,a}\, log_2 \frac{f_{msa,a}}{f_{bg,a}}$$

where the sum is over all 20 amino acids, $f_{msa,a}$ is the frequency of the amino acid a at a particular position in the MSA, and $f_{bg,a}$ is the background amino acid frequency of amino acid a in the whole MSA. If $f_{msa,a}$ equals to zero, this type of amino acid will be excluded from the calculation to avoid calculation error.

Bioinformatics analysis. FoldX prediction of $MLL3_{SET}$ was conducted by FoldX 5.0 (14) in the command line in the Windows operation system, using the crystal structure of $MLL3_{SET}$ (PDB ID:5f59) with the ligand removed. The missing loop was fixed by MODELLER (15) in UCSF Chimera. The fixed structure first went through the FoldX command RepairPDB to deal with bad torsion angles or Van der Waals' clashes. Then an individual mutation list was generated by a python script. Finally, FoldX command *BuildModel* was used to generate mutant models with the *numberOfRuns* parament set to 10.Anaconda distribution of Python 3.7.3 with Spyder integrated development environment was used to program all the python scripts in this work. A typical analysis of 919,626 sequences from the P_1 library running on a single Intel Core i7-8700K at 4.7 GHz took about 30 min to finish.

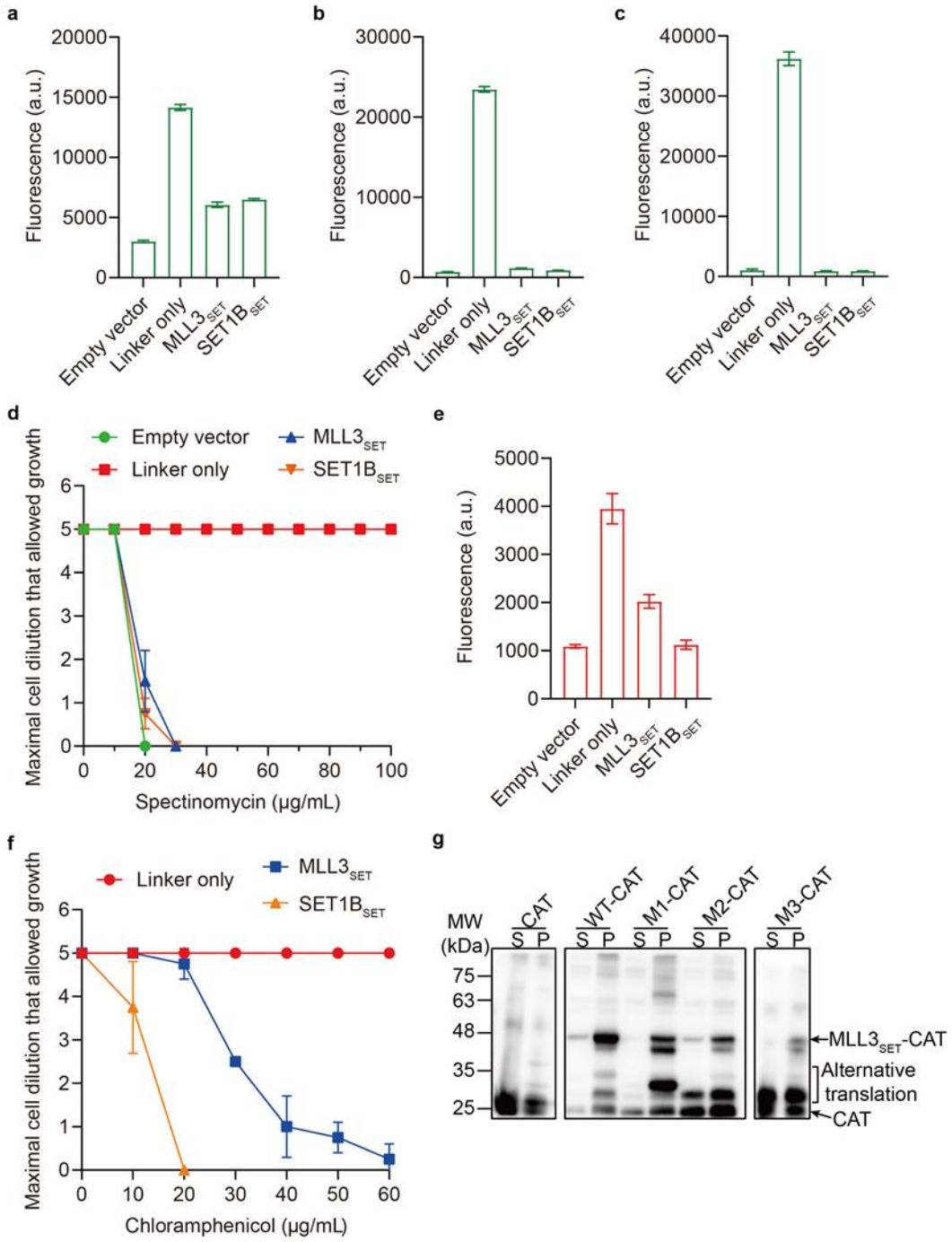

Fig. S1. Trial of different protein stability biosensors for detecting and evolving protein stability *in vivo*. a-c, Fluorescence of cells expressing a tripartite GFP folding reporter inserted with different POIs ($MLL3_{SET}$ and $SET1B_{SET}$). Expression of fusion proteins was induced at 16 ℃ (**a**), 30 ℃ (**b**), and 37 ℃ (**c**) respectively. **d**, Maximal cell dilution that allowed growth plotted against spectinomycin concentration. Cells were transformed with an aminoglycoside 3"-adenylyltransferase (ANT) biosensor inserted

with different POIs. **e**, Fluorescence of cells expressing the $CysG^{A}$ biosensor (developed in this study) inserted with different POIs. Expression of fusion proteins was induced at 30 ℃. **f**, Maximal cell dilution that allowed growth plotted against chloramphenicol concentration. Cells were transformed with a chloramphenicol acetyltransferase (CAT) biosensor fused with different POIs. The difference between $MLL3_{SET}$ and $SET1B_{SET}$-containing cells drove us to conduct a directed evolution experiment to isolate stability-enhanced $MLL3_{SET}$ mutants by selecting cells with increased chloramphenicol resistance. **g**, Western blot of various $MLL3_{SET}$-CAT fusion proteins expressed in chloramphenicol-resistant strains isolated from the directed evolution experiment. Soluble (S) and Precipitation (P) fractions of the cell lysates expressing these fusion proteins were analyzed by immunoblotting with antibodies against CAT. Mutations in $MLL3_{SET}$ do not increase the soluble fraction of their respective fusion proteins, but allow alternative translation of the fusion proteins and the formation of intact CAT due to degradation of the $MLL3_{SET}$ mutants. Therefore, we concluded that these $MLL3_{SET}$ mutants are false positives.

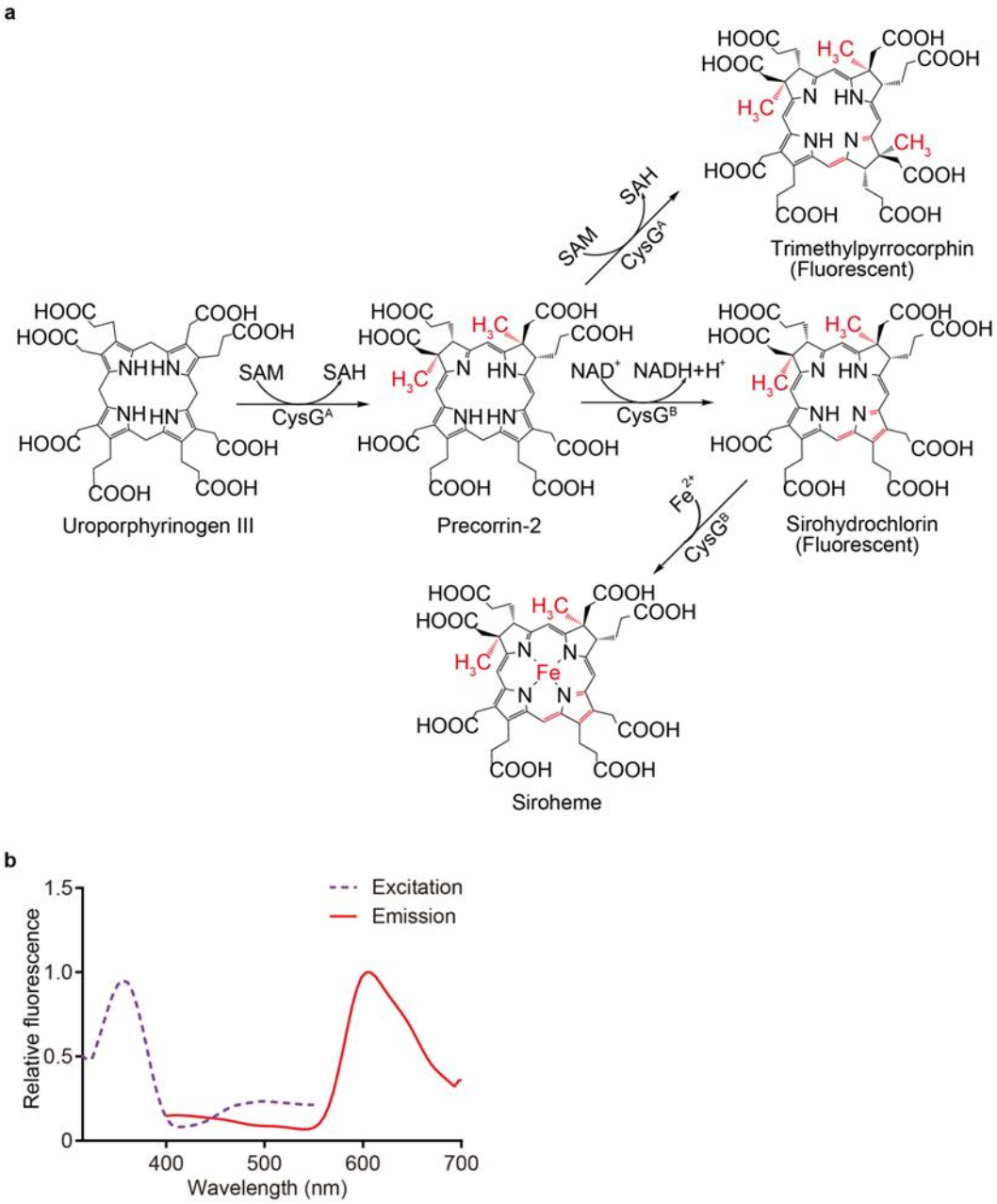

Fig. S2. The fluorescent phenotype conferred by CysG^{A}. a, Siroheme biosynthesis from uroporphyrinogen III in *E. coli*. Overexpressing of CysG^{A} results in the accumulation of the fluorescent compounds (trimethylpyrrocorphin and sirohydrochlorin). **b,** The fluorescent spectra of the extracts from cells expressing CysG^{A}. Excitation spectrum was recorded at an emission wavelength of 620 nm. Emission spectrum was recorded at an excitation wavelength of 357 nm.

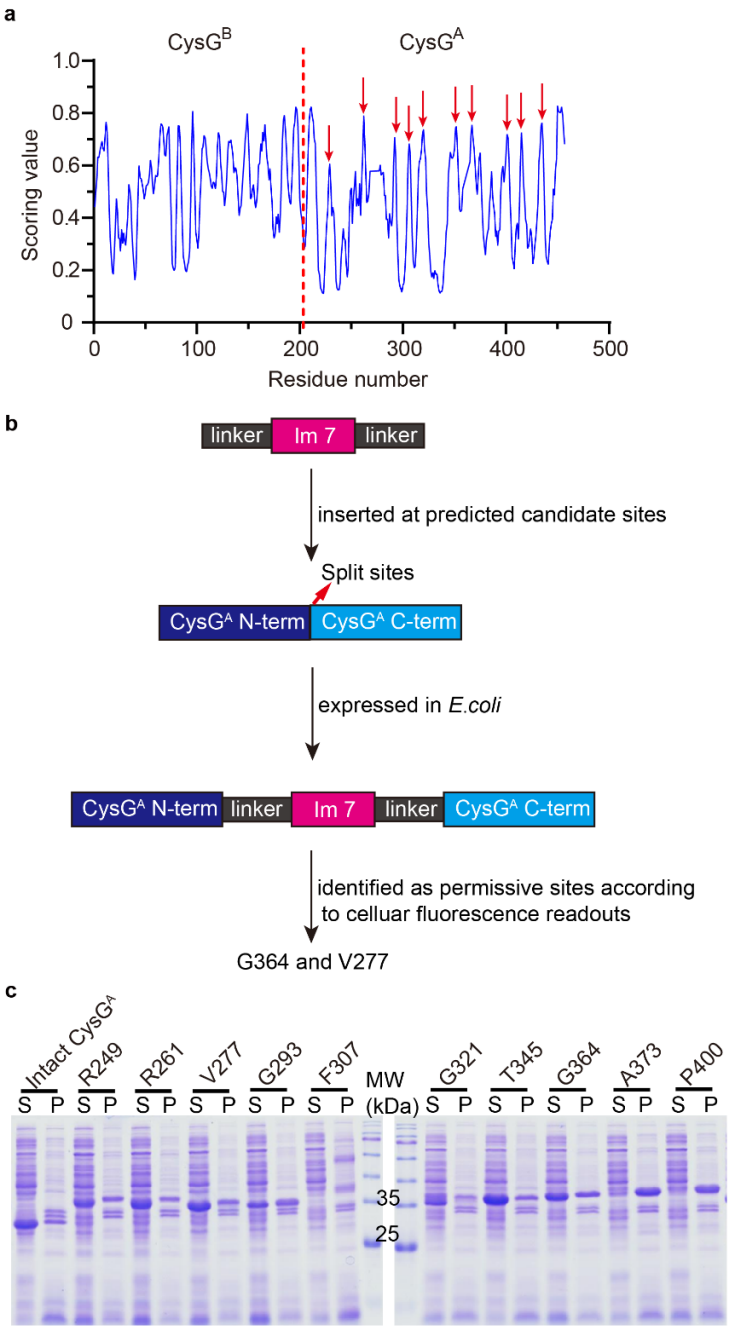

Fig. S3. Identification of permissive insertion sites in CysG^A. a, Prediction of circular permutation sites for CysG^A from *Salmonella typhimurium* through CPred. Sites with high scoring values may be permissive for the insertion of a foreign protein. Ten sites with top scoring values are indicated by red arrows and selected as candidates for insertion sites. **b,** Workflow of permissive sites identification. **c,** SDS-PAGE analysis of soluble and insoluble fractions of cells expressing CysG^A inserted with Im7 at different sites. S: soluble. P: pellet.

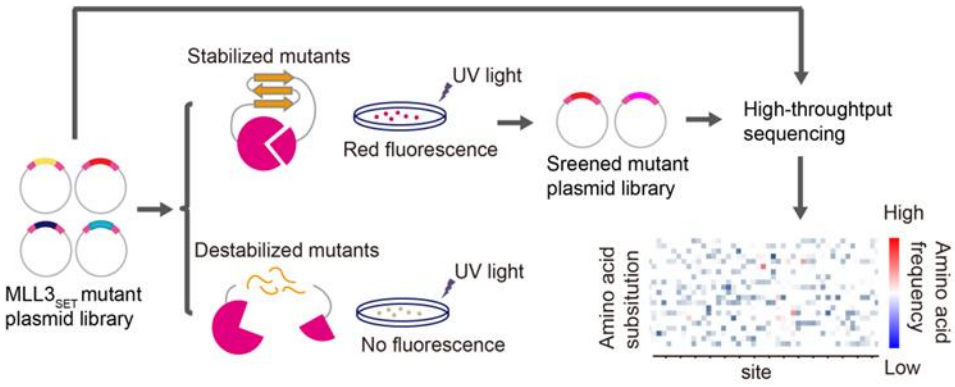

Fig. S4. Schematic diagram of deep mutational scanning of the $MLL3_{SET}$ protein. Individual members of the $MLL3_{SET}$ mutant library (the P_0 library) within the $CysG^A$ biosensor were screened by detecting fluorescence intensity. Colonies expressing stabilized variants in tripartite fusion emitted strong red fluorescence under UV light, whereas those containing destabilized variants had no red fluorescence or weak red fluorescence. The pools of unscreened and screened variants were then analyzed by high-throughput sequencing to obtain the frequency of amino acid substitutions at each site.

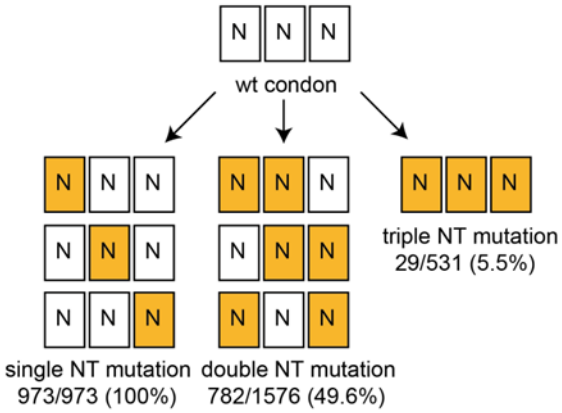

Fig. S5. The mutational coverage of MLL3$_{SET}$ full library (the P$_0$ library). There are 154 amino acid residues in the MLL3$_{SET}$ protein, so theoretically there are 3,080 (154 × 20) possible single-point amino acid substitutions including nonsense mutations. Among the 3,080 single-point substitutions, 973 can be achieved by only one nucleotide (NT) substitution, 1,576 require two nucleotide substitutions, and the remaining 531 need all three nucleotides substitutions. In our P$_0$ library, we have obtained all (973/973) amino acid substitutions that can be achieved by single nucleotide substitutions. The library also covers 49.6% (782/1,576) and 5.5% (29/531) of the amino acid substitutions that require two and three nucleotide substitutions, respectively. In total, the P$_0$ library contains 1,784 amino acid substitutions, which samples 57.9% of the possible substitutions for each amino acid at every position.

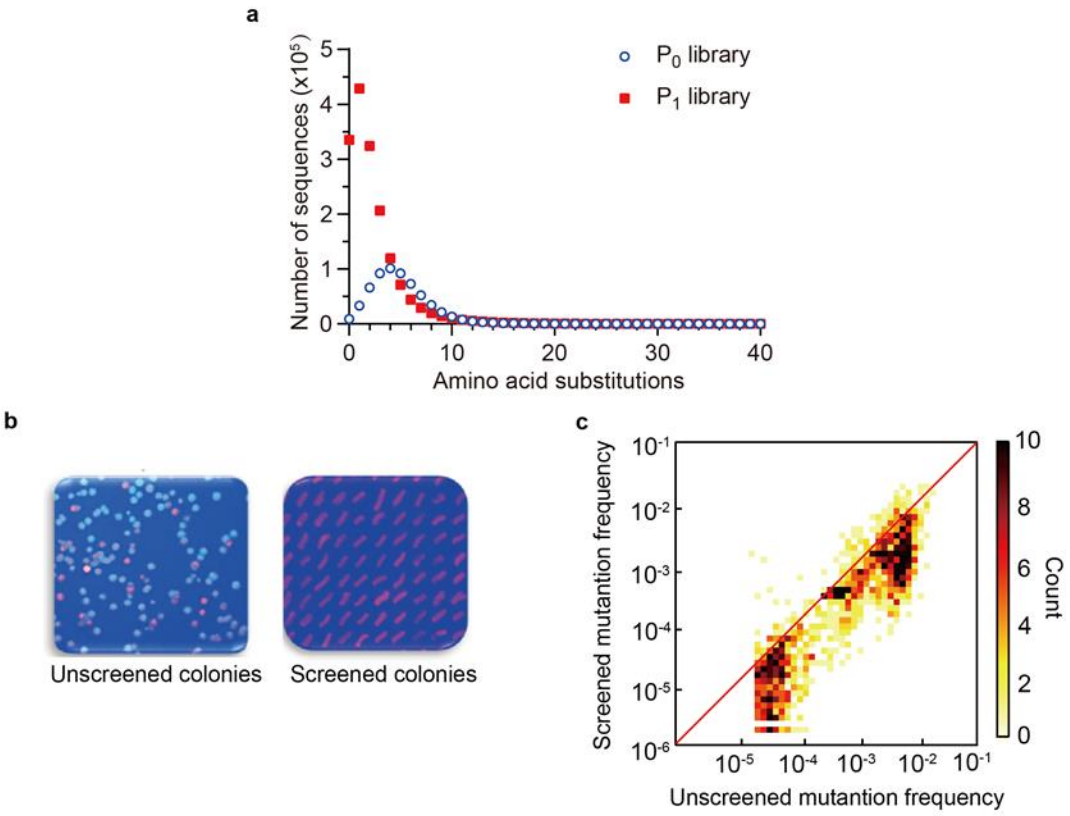

Fig. S6. Characterization of colonies and libraries before and after screening. a, Distribution of numbers of amino acid mutations per variant in the P_0 and P_1 libraries. In the P_0 library, numbers of substitutions per variant roughly follow a Poisson distribution with an average of 5 substitutions per sequence. After screening, an average of 2 substitutions per sequence in the P_1 library was obtained. **b,** Representative images of the colonies expressing the P_0 and P_1 library under UV light. **c,** The frequency of 1,784 amino acid substitutions in P_0 and P_1 libraries. Most substitutions decreased in frequency after the screening, and only a small number increased frequency under stability screening pressure.

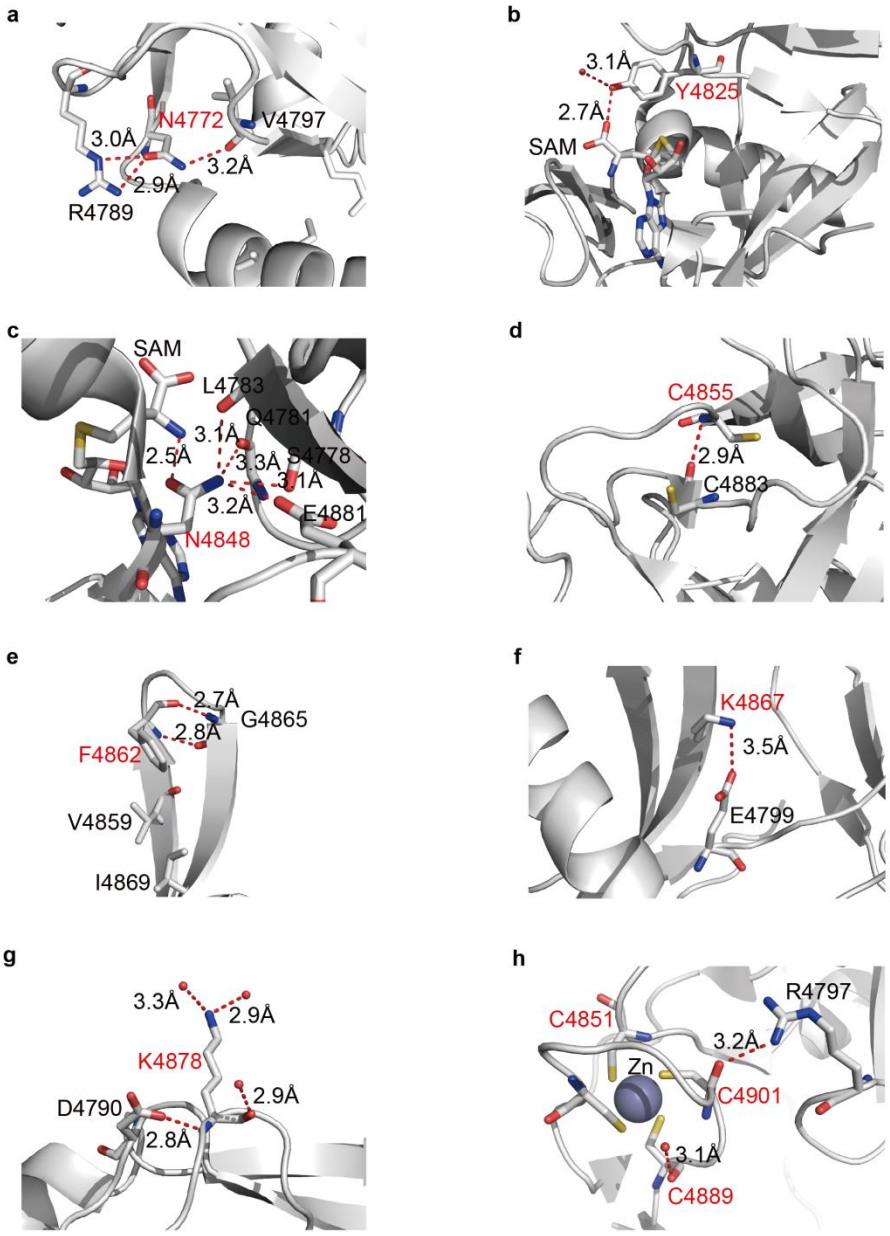

Fig. S7. Interaction networks of wild-type residues at the 10 top-ranking sites. a, N4772 forms hydrogen bonds with nearby R4789 and V4797. **b,** Y4825 forms hydrogen bonds with environmental water and cofactor SAM. **c,** N4848 involves in the hydrogen bond network with L4783, E4881, Q4781, and S4778. **d,** C4855 forms a hydrogen bond with nearby C4883. **e,** F4862 forms hydrogen bonds with G4865 and involves in hydrophobic contacts with nearby hydrophobic residues. **f,** K4867 forms a salt bridge with E4799. **g,** K4878 forms hydrogen bonds with environmental water and D4790. **h,** C4851, C4889, and C4901 are responsible for Zn^{2+} binding. C4889 also forms hydrogen

bonds with environmental water and C4901 forms a hydrogen bond with R4797. The distance of residues was calculated based on the $MLL3_{SET}$ crystal structure (PDB: 5F59).

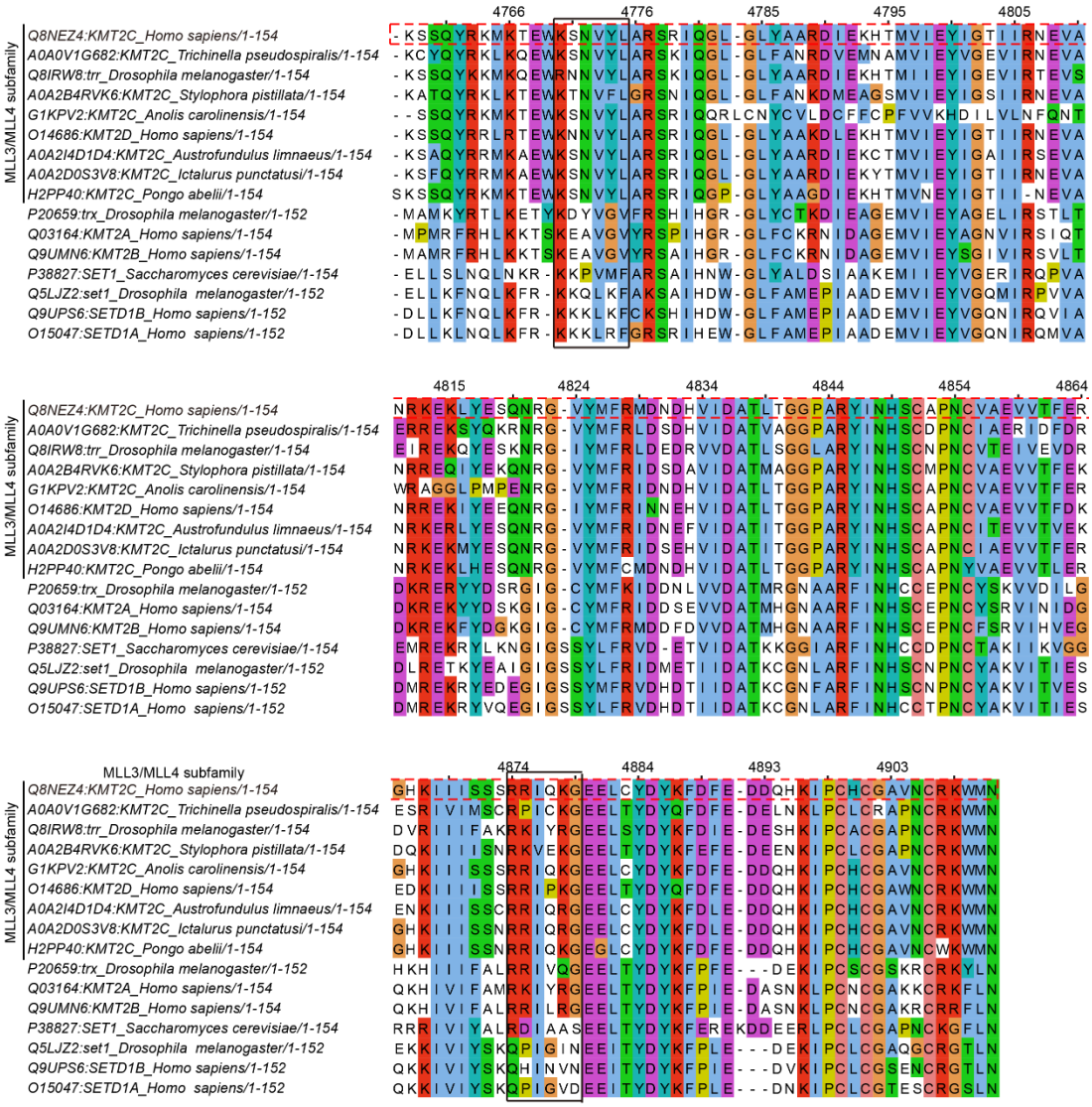

Fig. S8. Alignment of SET domain sequences in TRX/MLL family. A multiple sequence alignment was performed in MUSCLE 3.8.31 (15). The sequence position is annotated according to the MLL3 protein, whose SET domain sequence is outlined by red dashed lines. The black boxes mark the regions of K4770-L4775 and R4874-G4879. These regions are highly conserved among members of the MLL3/MLL4 subfamily. However, residues in this region are diverse in the whole TRX/MLL family.

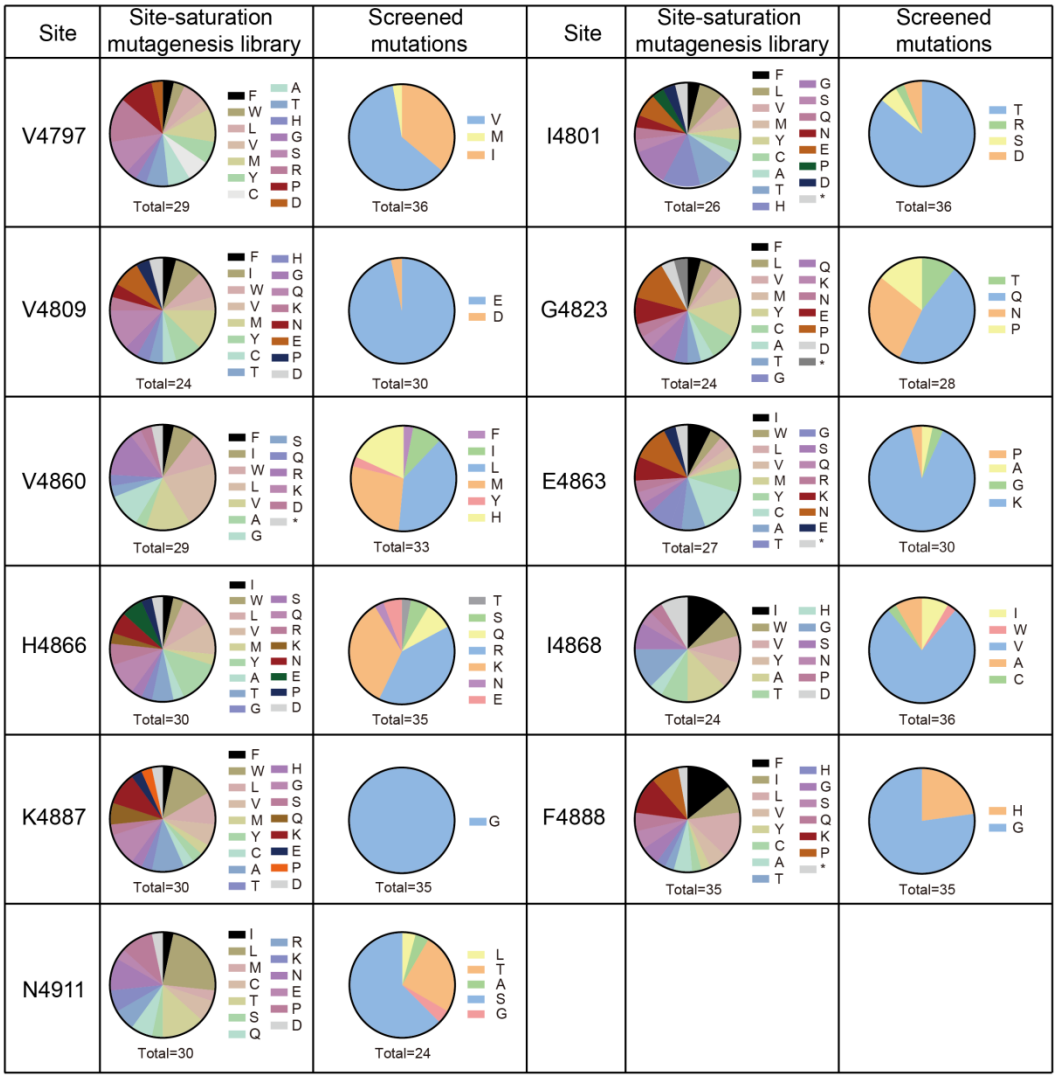

Fig. S9. The distribution of amino acid substitutions in site-saturation mutagenesis libraries and screened variants. According to the enrichment score, 11 top-ranking sites were selected to construct site-saturation mutagenesis libraries, which were achieved substituting with an NNK codon at the specific sites. All the site-saturation mutagenesis libraries were screened in the CysG^{A} system. Colonies with strong fluorescence were picked out, and the sequences of each construct were confirmed by Sanger sequencing. The unscreened libraries were also sequenced to confirm that saturation mutagenesis was achieved at each site.

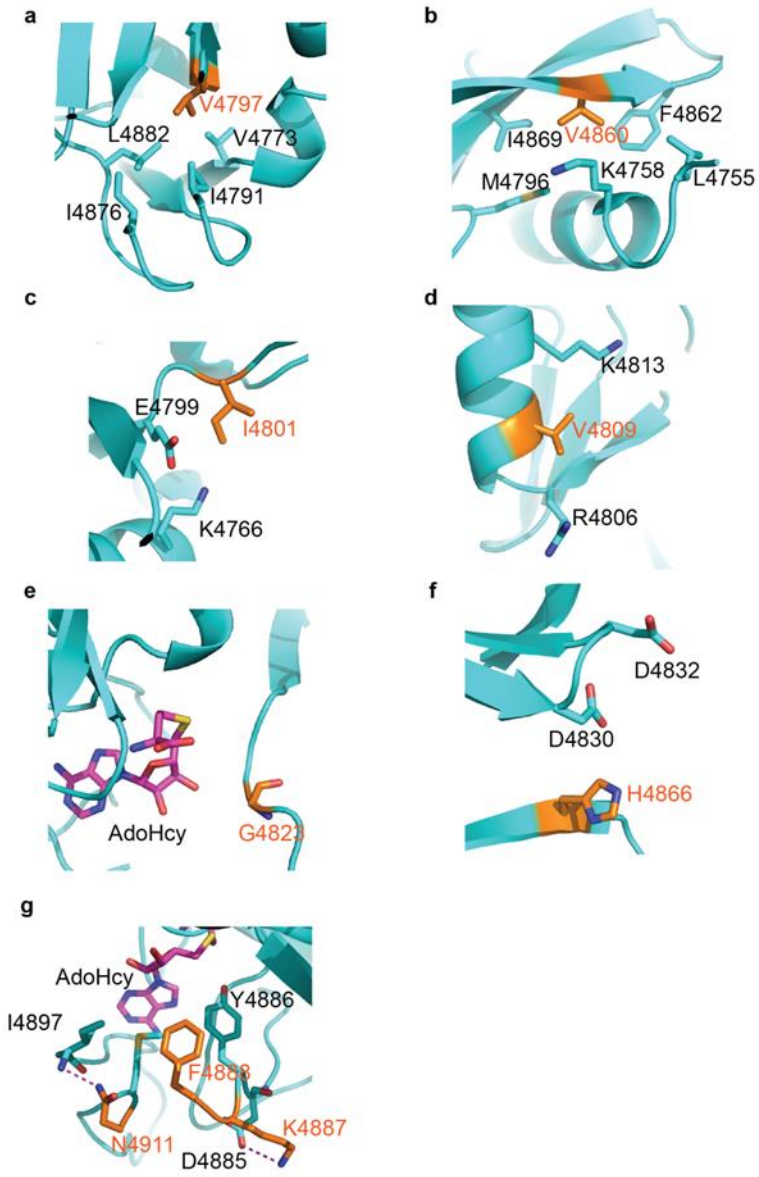

Fig. S10. Detailed views of the stabilization-hotspot residues in the MLL3$_{SET}$ structure.

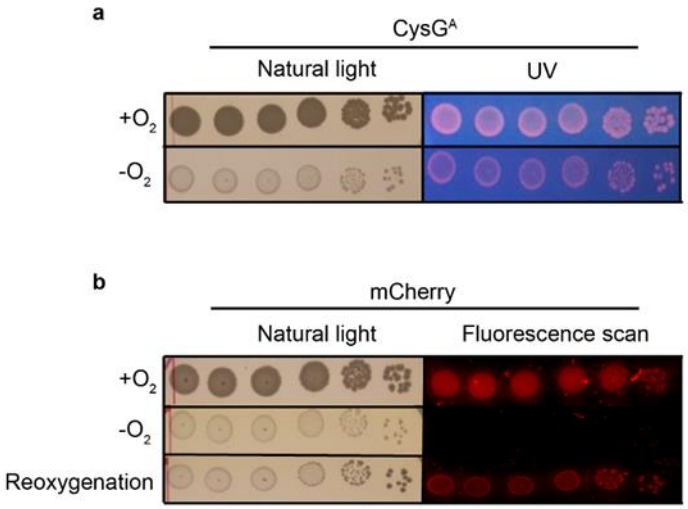

Fig. S11. The fluorescence readout of the CysG^A protein under anaerobic condition. Images of cells expressing CysG^A (**a**) or mCherry (**b**) grown under normal condition and hypoxia. In normal condition (aerobic), cells expressing CysG^A (**a**) or mCherry (**b**) exhibit strong fluorescence upon proper excitation. When grown in an anaerobic pouch (hypoxia), cells expressing CysG^A still exhibit strong fluorescence under UV light despite weaker growth (**a**). In contrast, cells expressing mCherry showed no fluorescence after hypoxia treatment. The fluorescence readouts of mCherry can be restored after a reoxygenation period of 6 hours (**b**).

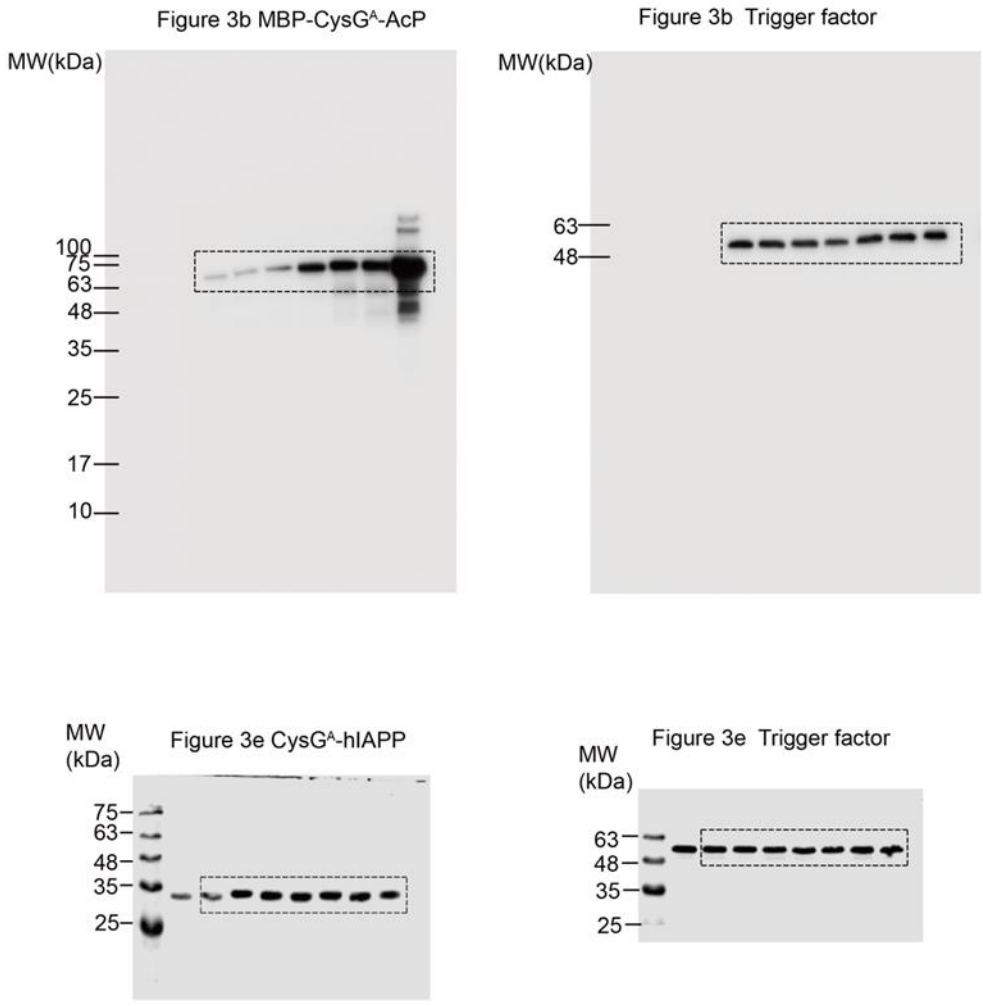

Fig. S12. Uncropped western blots related to Figure 3.

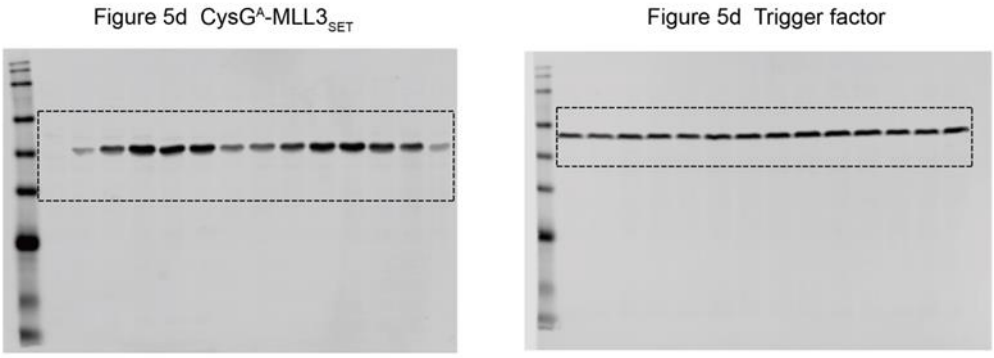

Fig. S13. Uncropped western blots related to Figure 5.

Table S1. Thermodynamic stabilities of test protein variants.

Im7 Mutant [a]	$\Delta\Delta G^{\circ}_{UN}$ (kJ/mol)	MBP Mutant [b]	$\Delta\Delta G^{\circ}_{UN}$ (kJ/mol)	AcP Mutant [c]	$\Delta\Delta G^{\circ}_{UN}$ (kJ/mol)
N26K T30N S58R	-5.89	T345I	-2.93	Y11F	-1.80
T30N	-2.31	Wild-type	0	Wild-type	0
Wild-type	0	V8G	4.60	V20A	1.20
I22V	8.90	A276G	6.28	E83D	6.30
I54V	11.00	G19C	9.62	M61A	16.60
L53A I54A	12.30	Y283D	13.39	L65V	22.40
F84A	22.00	G32D I33P	16.74		

Thermodynamic stabilities of test protein variants were given as $\Delta\Delta G^{\circ}_{UN}$, whereas $\Delta\Delta G^{\circ}_{UN} = \Delta G^{\circ}_{UN}$ (mutant) – ΔG°_{UN} (wild-type).

[a] Values from ref. 16.

[b] Values from ref. 17 and 18.

[c] Values from ref. 19. All AcP mutants in this paper contain the C21S substitution, but they will be referred to as single point mutants.

Table S2. Energy terms of FoldX predictions of 11 stabilizing mutation of $MLL3_{SET}$.

	H-bond	Van der Waals	Electrostatics	Solvation polar	Solvation hydrophobic	Entropy
V4797I	-0.075	-0.495	0.000	+0.184	-1.061	+0.588
I4801T	+0.064	0.607	+0.120	-0.777	+1.087	-0.536
V4809E	-1.540	-0.135	+0.032	+0.500	+0.036	+0.748
G4823Q	-1.173	-0.566	-0.004	+0.726	-0.678	+0.810
V4860M	-0.034	-0.374	-0.084	+0.185	-1.370	+1.765
E4863K	0.000	+0.005	+0.430	-0.022	-0.001	-0.157
H4866R	+0.477	-0.167	-1.112	+0.282	-0.265	-0.011
I4868V	+0.022	+0.579	-0.010	-0.324	+1.167	-0.562
K4887G	+1.195	+0.251	+0.933	-2.854	+1.435	-1.861
F4888G	+0.003	+1.139	-0.031	-1.127	+2.233	-0.886
N4911S	+1.290	+0.574	-0.049	-0.739	0.975	-1.391

FoldX prediction procedure is included in the Method part. Each the energy term is defined by the equation, $\Delta\Delta G°_{UN} = \Delta G°_{UN}$ (mutant) – $\Delta G°_{UN}$ (wild-type) in kcal/mol. Solvation polar and solvation hydrophobic are the differences in solvation energy for polar and polar groups respectively when these groups switch from the unfolded to the folded state. Entropy depict the cost of fixing the backbone and side chain in the folded state. The apparently stabilizing terms for each mutation are highlighted in yellow.

Table S3. Plasmids used in this work.

Plasmid	Relevant characteristics	Source
pTrc-99a	Expression vector	Pharmacia
pTrc-CysG^{A}	*cysG*^{A} cloned into pTrc99a derived vector pssTrx	This study
pTrc-CysG^{A}-R249	Im7 WT inserted into *cysG*^{A} at R249 via GS linker	This study
pTrc-CysG^{A}-R262	Im7 WT inserted into *cysG*^{A} at R261 via GS linker	This study
pTrc-CysG^{A}-V277	Im7 WT inserted into *cysG*^{A} at V277 via GS linker	This study
pTrc-CysG^{A}-G293	Im7 WT inserted into *cysG*^{A} at G293 via GS linker	This study
pTrc-CysG^{A}-F307	Im7 WT inserted into *cysG*^{A} at F307 via GS linker	This study
pTrc-CysG^{A}-G321	Im7 WT inserted into *cysG*^{A} at G321 via GS linker	This study
pTrc-CysG^{A}-T345	Im7 WT inserted into *cysG*^{A} at T345 via GS linker	This study
pTrc-CysG^{A}-G364	Im7 WT inserted into *cysG*^{A} at G364 via GS linker	This study
pTrc-CysG^{A}-A373	Im7 WT inserted into *cysG*^{A} at A373 via GS linker	This study
pTrc-CysG^{A}-P400	Im7 WT inserted into *cysG*^{A} at P400 via GS linker	This study
pTrc-CysG^{A}-IM7 variants	IM7 varaints inserted into *cysG*^{A} at G364 via GS linker	This study

pTrc-CysG^{A}-MBP variants	MBP varaints inserted into *cysG*^{A} at G364 via GS linker	This study
pTrc-MBP-CysG^{A}-AcP variants	AcP varaints inserted into *cysG*^{A} at G364 via GS linker containing an N-terminal MBP tag	This study
pTrc-CysG^{A}-PolyQ tracts	PolyQ tracts with different lengths inserted into *cysG*^{A} at G364 via GS linker	This study
pET28b-His-SUMO-MLL3$_{SET}$	For MLL3$_{SET}$ purification	Ref. 5
pET28b-His-SUMO-CysG^{A}	For CysG^{A} purification	This study
pET28b-His-SUMO-AcP variants	For AcP variants purification	This study
pTrc-CysG^{A}-MLL3$_{SET}$ variants	MLL3$_{SET}$ variants inserted into *cysG*^{A} at G364 via GS linker	This study
pTrc-mCherry	For mCherry overexpression	This study

SI References

1. A. Bryksin, I. Matsumura, Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques* **48**, 463-465 (2010).
2. M. J. Warren *et al.*, Gene dissection demonstrates that the Escherichia coli cysG gene encodes a multifunctional protein. *Biochem. J.* **302**, 837-844 (1994).
3. S. Quan *et al.*, Genetic selection designed to stabilize proteins uncovers a chaperone called Spy. *Nat. Struct. Mol. Biol.* **18**, 262-269 (2011).
4. A. Malik, A. Mueller-Schickert, J. C. A. Bardwell, Cytosolic Selection Systems To Study Protein Stability. *J. Bacteriol.* **196**, 4333-4343 (2014).
5. Y. Li *et al.*, Structural basis for activity regulation of MLL family methyltransferases. *Nature* **530**, 447-452 (2016).
6. M. Martin, CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10-12 (2011).
7. A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, J. D. Neufeld, PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
8. F. Chiti *et al.*, Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Mol. Biol.* **6**, 1005-1009 (1999).
9. M. M. Santoro, D. W. Bolen, Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* **27**, 8063–8068 (1988).
10. C. Silva, M. Martins, S. Jing, J. Fu, A. Cavaco-Paulo, Practical insights on enzyme stabilization. *Crit. Rev. Biotechnol.* **38**, 335-350 (2018).
11. M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, C. O. Wilke, Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**, e80635 (2013).
12. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
13. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
14. J. Delgado, L. G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).
15. A. Fiser, R. K. Do, A. Sali, Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).
16. A. P. Capaldi, C. Kleanthous, S. E. Radford, Im7 folding mechanism: misfolding on a path to the native state. *Nat. Struct. Biol.* **9**, 209-216 (2002).
17. S. Y. Chun, Strobel, S., P. Bassford, Jr, L. L. Randall, Folding of maltose-binding protein. Evidence for the identity of the rate-determining step in vivo and in vitro. *J. Biol. Chem.* **268**, 20855–20862 (1993).
18. J. M. Betton, and M. Hofnung, Folding of a mutant maltose-binding protein of Escherichia coli which forms inclusion bodies. *J. Biol. Chem.* 271, 8046-8052 (1996)

19. F. Chiti *et al.*, Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Mol. Biol.* **6**, 1005-1009 (1999).