Expanded chemical diversity sampling through whole protein evolution.

Amy J. Baldwin[a,b], James A. J. Arpino[a], Wayne R. Edwards[a,b], Eric M. Tippmann[b]*, D. Dafydd Jones[a]*

[a] School of Biosciences and [b] School of Chemistry, Main Building, Park Place, Cardiff University, Cardiff CF10 3AT, UK.

Supporting Information

Supporting Methods

Materials

The engineered transposon, termed MuDel, used to delete contiguous trinucleotide sequences at random positions in a target gene was constructed as described previously [1]. Previous studies showed that MuDel insertion into target DNA was essentially random [1-3]. The DNA cassette used to donate the TAG trinucleotide sequence to the target gene, termed SubSeq^{TAG}, was synthesised in the same manner as the SubSeq^{NNN} constructed as described previously [4] with the exception that primer AJBtag025 (5'-[Phos]TAGGAAAGGACTCAGTGTGTCGGCGGCCGCGGATCCT-3') replaced DDJdi041. The replacement TAG trinucleotide sequence (underlined above) was situated at the 5' end of AJBtag025, which replaced the 3 randomised nucleotides of the original SubSeq^{NNN} cassette.

The plasmid, pNOM, containing the *bla* gene used as one of the targets was constructed as described previously [1]. The gene (*eGFP*) encoding enhanced green fluorescent protein (eGFP) from plasmid pEGFP-N3 (Clontech) was amplified by PCR with primers AJBgfp009 (5'-AGCAGACCATATGGTGAGCAAGGGCGAGGAGC-3') and AJBgfp010 (5'-

ATACTCGAGTTACTTGTACAGCTCGTCCATGCCG-3'). The PCR product was subsequently digested with NdeI and XhoI (NE Biolabs) and cloned between the NdeI and XhoI sites of an in-house expression plasmid, containing an engineered T7 promoter sequence and the ampicillin (Amp) resistance gene, to produce the vector pNOM-XP3-*eGFP*.

DNA was purified using the following kits supplied by Qiagen: Qiagen Plasmid Midi kit (from cell culture), QIAprep Spin Miniprep kit (from cell culture), a QIAquick Gel Extraction kit (from agarose gels) or QIAquick PCR purification kit (from PCR reaction mixtures).

Construction of the *bla*^{TAG} and *eGFP*^{TAG} libraries.

1. Construction of the BLAΔ trinucleotide deletion libraries

A library containing 1000 different clones with the MuDel transposon inserted randomly within the *bla* gene was a subset of a larger library termed BLAΔ^{1664} constructed previously [4]. This library was termed BLAΔ^{1000}. The estimated size of the BLAΔ^{1000} library was deemed sufficient to cover all potential MuDel insertion positions within *bla* (coding region of 861 bp), even when redundancy in transposon insertion positions (10-15%) was factored in [1,3].

2. Construction of the eGFPΔ trinucleotide deletion libraries

A generic approach not reliant on the functional properties of the target protein was taken for constructing the library constituting the MuDel transposon inserted randomly within *eGFP* (outlined in Supporting Figure 2). Insertion of MuDel into the target DNA was performed by *in vitro* transposition with the HyperMu™ MuA transposase system (Epicentre® Biotechnologies) using 1 unit of the transposase, 300 ng of pNOM-XP3-*eGFP* and 40 ng of MuDel. The reaction mixture was incubated at 37°C for 3 hr and transposition terminated by the addition of 0.1% w/v SDS. Chemically competent *E. coli* NovaBlue GigaSingle cells were transformed with *ca* 50 ng DNA from the transposition reaction. A proportion of the transformed cells (1.5%) were grown on LB agar supplemented with 20 μg/ml chloramphenicol (Cam) to determine the number of clones with MuDel inserted within the whole plasmid (calculated to be a total of 7590). The proportion of clones containing MuDel within *eGFP* was determined by PCR with primers AJBgfp009 and AJBgfp010. From 48 clones analysed, 16 (33%) were deemed to contain MuDel inserted within *eGFP* (fragment size 2039 bp compared to 732 bp for *eGFP* alone). The rest of the transformed cells were used to inoculate 1l LB broth supplemented with 20 μg/ml Cam and incubated at 37°C overnight. Subsequently, DNA comprising MuDel randomly inserted throughout the plasmid was purified from the culture. To isolate the population of species with MuDel inserted within *eGFP*, 3 μg of the plasmid DNA was digested with NdeI and XhoI, and the products of digestion separated by agarose

gel electrophoresis. The fragments equivalent to the plasmid backbone (2143 bp) and *eGFP* containing MuDel (2030 bp) were isolated (Supporting Figure 2) and purified. The two purified fragments were then ligated using T4 DNA ligase followed by transformation of chemically competent *E. coli* NovaBlue GigaSingles. Transformation efficiency was determined to be 9×10^5 colony forming units (cfu)/μg DNA in the ligation reaction. Transformed cells were used to inoculate 1l of LB broth supplemented with 20 μg/ml Cam and cultures grown overnight at 37°C. Plasmid DNA with MuDel randomly inserted within *eGFP* was purified from the cultures. The library was calculated to contain 2504 individual clones as PCR analysis revealed that 33% of the original 7590 clones (see above) with plasmid-borne MuDel had the transposon inserted within *eGFP*. This library was termed eGFPΔ^{2504}, and restriction analysis confirmed that insertions were distributed randomly throughout *eGFP* (data not shown). The estimated sizes of the eGFPΔ^{2504} library was deemed sufficient to cover all potential MuDel insertion positions within *eGFP* (coding region 720 bp), even when redundancy in transposon insertion positions (10-15%) was factored in [1,3].

3. Construction of the TAG trinucleotide replacement libraries: *bla*^{TAG} and *eGFP*^{TAG}

Clones comprising either the BLAΔ^{1000} or eGFPΔ^{2504} library were pooled and plasmid DNA isolated. Plasmid DNA was digested with MlyI (NE Biolabs) for 150 min at 37°C and dephosphorylated using alkaline phosphatase. Linearised plasmid DNA was separated from the MuDel fragment by agarose gel electrophoresis and bands of the expected size were purified. The SubSeq^{TAG} DNA cassette was inserted within the cut plasmid using T4 DNA ligase in a reaction that contained a total of 250 ng DNA in a 1:3 molar ratio (vector:SubSeq^{TAG}) for 30 min at room temperature. *E. coli* DH5α cells were transformed by electroporation with 1/10th of the ligation reaction, with a transformation efficiency of 3-4 × 10^4 cfu/μg DNA.

A total of 960 and 1877 individual clones (from a total of >3000) derived from the insertion of SubSeq^{TAG} into *bla* or *eGFP*, respectively, that displayed resistance to kanamycin (Kan) were picked at random and used to inoculate separate wells of 96-well microtitre plates containing 100 μl LB broth and 50 μg/ml Kan. Overnight cultures grown at 37°C were pooled and DNA was purified. SubSeq^{TAG} was removed from the plasmid by digestion of the plasmid pool with MlyI for 1 h at 37°C (*ca* 2 μg DNA per 50 μl reaction volume). MlyI was inactivated by incubation at 65°C for 20 min and linearised DNA purified after agarose gel electrophoresis. Plasmid (*ca* 100 ng) was recircularised using a T4 DNA ligase to reform the full-length target gene containing a randomly-placed TAG trinucleotide exchange. These trinucleotide exchange libraries based on either *bla* or *eGFP* were termed *bla*^{TAG} and *eGFP*^{TAG}, respectively.

Screening for TEM-1 variants that tolerate incorporation of *p*-iodophenylalanine

The engineered orthogonal $tRNA^{p\text{-iodoPhe}}_{CUA}$ /*p*-iodophenylalanine (*p*-iodoPhe) tRNA synthetase system described previously [5] was used to incorporate *p*-iodoPhe in response to the UAG codon instead of terminating translation at this point. The DNA fragments corresponding to the *Methanococcus jannaschii* tRNA and synthetase were synthesized with the reported mutations [6] (Epoch Biolabs Inc.) to produce a plasmid expressing the orthogonal pair (p6tRNA/Iodo). *E. coli* BL21-Gold(DE3) electrocompetent cells (Stratagene) containing p6tRNA/Iodo vector were transformed with *ca* 10 ng of bla^{TAG} library DNA derived from the self-ligation reaction used to reconstruct the full-length *bla* gene after SubSeq removal. The transformation efficiency was >6.0 × 10^5 cfu/μg DNA. Transformed cells were subsequently grown on LB agar containing 100 μg/ml Amp, 34 μg/ml Cam and 1 mM *p*-iodoPhe (Aldrich).

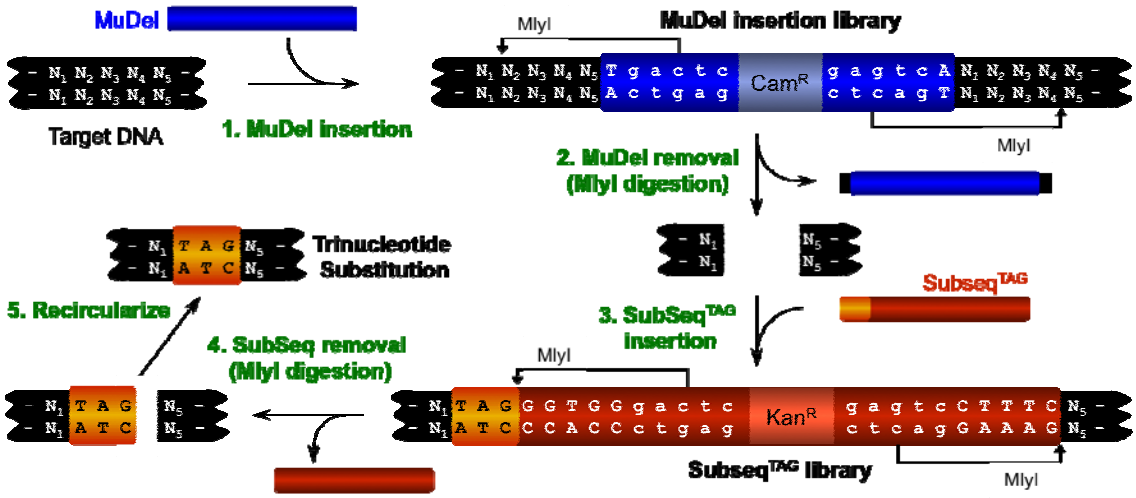
Variants incorporating the unnatural amino acid were identified in a positive-negative *in vivo* selection process. A total of 456 discrete colonies growing in the presence of *p*-iodoPhe were arbitrarily chosen and used to individually inoculate 100 μl 96-well microtitre plates containing LB broth supplemented with 34 μg/ml Cam. After a brief period of shaking at 37°C, cells were transferred in duplicate to LB agar plates (supplemented with 100 μg/ml carbenicillin (Carb) and 34 μg/ml Cam) in the presence and absence of 1 mM *p*-iodoPhe using a 96-pronged replicating fork and incubated at 37°C for 24 h. Of the 456 clones tested, 57 were unable to grow in the absence of 1 mM *p*-iodoPhe and were deemed to require the incorporation of the unnatural amino acid to produce functional TEM-1, and hence contained an in-frame TAG. After ligation, the expected proportion of in-frame TAG codons in the bla^{TAG} library was 1 in 6 or 16.7% (2 orientations and 3 frames) assuming insertions are distributed entirely randomly throughout the gene. However, it is unlikely that every clone contained a mutation that produced functional protein. Since we observed that 1 in 8 clones (12.5%) were sensitive to Carb in the absence of *p*-iodoPhe it follows that many residues in TEM-1 may not have tolerated substitution with *p*-iodoPhe, as would be expected. The *bla* gene was isolated from clones that regained antibiotic resistance in the presence of *p*-iodoPhe by PCR with primers flanking the gene (5'-TCCGCTCATGAGACAATAACCCTG-3' and 5'-CTACGGGGTCTGACGCTCAGTG-3'). The PCR product was sequenced (DNA Sequencing Core, Molecular Biology Unit, Cardiff University) using the same primers to determine the position and nature of the mutation.

Screening for eGFP variants that tolerate incorporation of *p*-iodophenylalanine

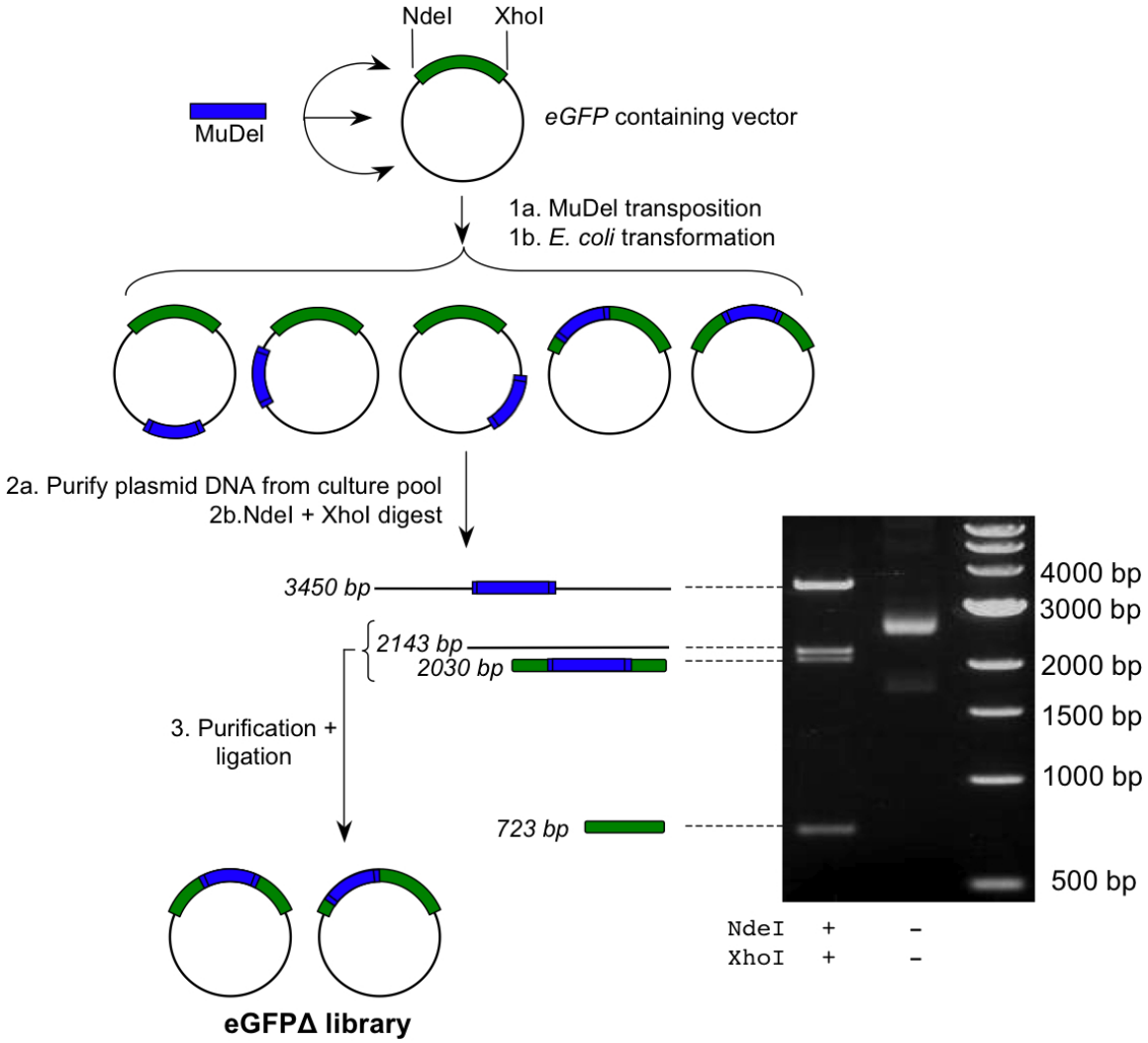
E. coli BL21-Gold(DE3) electrocompetent cells were transformed with *ca* 10 ng of $eGFP^{TAG}$ library DNA derived from the self-ligation reaction used to reconstruct the full-length *eGFP* gene after SubSeq removal. Variants incorporating the unnatural amino acid were identified in a negative-

positive *in vivo* selection process. The transformed cells were grown in the absence of *p*-iodoPhe on LB agar containing 100 μg/ml Amp and 10 μM IPTG, and incubated at 37°C for 16 h, followed by a further 24 h at 4°C to allow eGFP *in vivo* fluorescence to develop. Colonies were screened visually for green colouration after excitation in the near UV using a UV transilluminator to identify clones producing mature eGFP. Clones containing functional eGFP in the absence of *p*-iodoPhe were assumed not to contain an in-frame TAG that would otherwise prematurely truncate the protein; these clones were estimated to comprise *ca* 60% of all the clones. From a total of *ca* 10,000 colonies, 846 non-fluorescent colonies were arbitrary chosen and grown individually in 100 μl of LB broth containing 200 μg/ml Amp at 37°C for 16 h in 96-well microtitre plates. The individual liquid cultures were pooled and plasmid DNA isolated to create the non-fluorescent sub-library of the *eGFP*^{TAG} library enriched with in-frame TAG sequences. Electrocompetent *E. coli* BL21-Gold(DE3) containing p6tRNA/Iodo were transformed with 20 ng of *eGFP*^{TAG} sub-library DNA (efficiency of >1.4 × 10^7 c.f.u./μg library DNA) and grown on LB agar (supplemented with 34 μg/ml Cam, 200 μg/ml Carb and 10 μM IPTG) in the absence and presence of 1 mM *p*-iodoPhe at 30°C for 48 h. Colonies that displayed a green colour after excitation in the near UV comprised 18% of the total colonies (sample of 478 scored) and were assumed to contain an in-frame TAG.

The *eGFP* gene was isolated from 26 different clones that regained fluorescence in the presence of *p*-iodoPhe by PCR with primers flanking the gene (5'-TGCTCACATGTGCGTAGAGG-3' and 5'-CAGGGTTATTGTCTCATGAGCGGA-3'). The PCR product was sequenced (DNA Sequencing Core, Molecular Biology Unit, Cardiff University) using the same primers to determine the position and nature of the mutation.



Supporting Figure 1. Mechanism for generating TAG libraries. (1). Insertion of the engineered transposon MuDel (blue) results in the duplication of 5 bp ($N_1N_2N_3N_4N_5$) of the target gene at the insertion point. Two MlyI recognition sites (5' GAGTC$(N)_5$ 3') were placed 1 bp away from the site of transposon insertion. MlyI cuts 5 bp outside its recognition sequence to generate a blunt end. (2). MuDel is removed by MlyI digestion along with 4 bp of target DNA at each end (8 bp in total), which equates to deletion of a 3 bp contiguous sequence ($N_2N_3N_4$) from the target gene. (3) SubSeq^{TAG} (red and yellow) is inserted into the break in the target gene after MuDel removal. Two MlyI recognition sites are strategically placed towards the termini of the cassette; one is located so that digestion at the point where SubSeq^{TAG} joins the target DNA and the second will cut 3 bp into SubSeq^{TAG} so donating 3 bp (TAG, yellow) to the target DNA. (4). Removal of SubSeq by digestion with MlyI results in the replacement of the 3 bp deleted from the target gene with TAG. (5) Intramolecular ligation reforms the target but with one contiguous trinucleotide sequence replaced with TAG.



Supporting Figure 2. Isolation of clones containing MuDel inserted within *eGFP*. After transposition (1a) with MuDel (blue) into vector containing e*GFP* and transformation of *E. coli* (1b), plasmid DNA was purified from a culture of pooled clones displaying chloramphenicol resistance (2a). The plasmid pool was digested with NdeI and XhoI (2b) to produce the observed digestion pattern. The size of the fragment depends on the location of MuDel within the vector as outlined on the diagram. Fragments with the equivalent size to the plasmid backbone without *eGFP* (black line; 2143 bp) and the *eGFP* gene with MuDel inserted within it (green and blue; 2030 bp). These two fragments were purified together and ligated (3a) to form the eGFPΔ library with MuDel randomly inserted within the *eGFP* region of the expression vector.

Supporting Table 1. Sequence variation and observed frequency of *bla* variants with *p*-iodoPhe-dependent TEM-1 *in vivo* activity.

***bla* variant**	**Nucleotides replaced [a]**	**Replacement trinucleotide**	***p*-iodoPhe insert position [b]**	**Frequency [c]**
bla1	^{73}CAA^{75}	TAG	P27	8
bla2	^{109}CAG^{111}	TAG	Q39	1
bla3	^{121}CGA^{123}	TAG	R43	1
bla4	^{139}CTG^{141}	TAG	L49	1
bla5	^{226}GGT GCG^{231}	CTA	$_{(G78G)}$A79	1
bla6	^{229}GCG^{231}	TAG	A79	2
bla7	^{256}CAA^{258}	TAG	Q88	2
bla8	^{286}TCT CAG^{291}	CTA	$_{(S98S)}$Q99	1
bla9	^{301}GTT GAG^{306}	CTA	$_{(V103V)}$E104	1
bla10	^{334}ACG^{336}	TAG	T114	1
bla11	^{379}ATG^{381}	TAG	M129	4
bla12	^{469}GTA^{471}	TAG	V159	1
bla13	^{505}GAA^{507}	TAG	E171	1
bla14	^{511}ATA^{513}	TAG	I173	8
bla15	^{622}TGG^{624}	TAG	W210	5
bla16	^{631}GCG^{633}	TAG	A213	1
bla17	^{649}CCA^{651}	TAG	P219	1
bla18	^{745}CCA^{747}	TAG	P252	1
bla19	^{760}TCC^{762}	TAG	S258	1
bla20	^{780}AGT CAG^{785}	CTA	$_{(S268S)}$Q269	10
bla21	^{817}AGA^{819}	TAG	R277	1
bla22	^{841}TCA^{843}	TAG	S285	3
bla23	^{844}CTG^{846}	TAG	L286	1

[a] Groups of three nucleotides represent a codon. When 6 nucleotides are shown, those underlined were replaced. [b] Predicted from location of the in-frame TAG in the *bla* gene. [c] Number of times variant observed.

Supporting Table 2. Sequence variation and observed frequency of *eGFP* variants with *p*-iodoPhe-dependent *in vivo* fluorescence.

***eGFP* variant**	**Nucleotides replaced [a]**	**Replacement trinucleotide**	***p*-iodoPhe insert position [b]**	**Frequency [c]**
eGFP1	^{1}ATG GTG^{6}	TAG	*Null (M1I)*	1
eGFP2	^{31}GGG^{33}	TAG	G10	3
eGFP3	^{46}CTG^{48}	TAG	L15	1
eGFP4	^{103}GAG^{105}	TAG	E34	1
eGFP5	^{226}CCC^{228}	TAG	P75	1
eGFP6	^{241}CAG^{243}	TAG	Q80	1
eGFP7	^{304}AAG^{306}	TAG	K101	1
eGFP8	^{322}AAG^{324}	TAG	K107	1
eGFP9	^{367}CGC^{369}	TAG	R122	1
eGFP10	^{394}AAG GAG^{399}	CTA	K131N, E132	1
eGFP11	^{412}CTG^{414}	TAG	L137	1
eGFP12	^{505}CGC^{507}	TAG	R168	8
eGFP13	^{604}CTG^{606}	TAG	L201	1
eGFP14	^{613}CAG^{615}	TAG	Q204	1
eGFP15	^{655}ATG GTC^{660}	TAG	*Null (M218I)*	1
eGFP16	^{664}CTG^{666}	TAG	L221	2

[a] Groups of three nucleotides represent a codon. When 6 nucleotides are shown, those underlined were replaced. [b] Predicted from location of the in-frame TAG in the *eGFP* gene. eGFP has an additional amino acid at the second position when compared with wild-type GFP. For consistency we refer to this additional second residue as 1*. Variants in italics do not contain an in-frame TAG, with resulting amino acid substitution shown in brakets. [c] Number of times variant observed.

Supporting References

(1) Jones, D. D. *Nucleic Acids Res.* **2005**, *33*, e80.

(2) Haapa, S.; Taira, S.; Heikkinen, E.; Savilahti, H. *Nucleic Acids Res.* **1999**, *27*, 2777-84.

(3) Simm, A. M.; Baldwin, A. J.; Busse, K.; Jones, D. D. *FEBS Lett.* **2007**, *581*, 3904-8.

(4) Baldwin, A. J.; Busse, K.; Simm, A. M.; Jones, D. D. *Nucleic Acids Res.* **2008**, *36*, e77.

(5) Xie, J.; Wang, L.; Wu, N.; Brock, A.; Spraggon, G.; Schultz, P. G. *Nat. Biotechnol.* **2004**, *22*, 1297-301.

(6) Ryu, Y.; Schultz, P. G. *Nat. Methods* **2006**, *3*, 263-5.