

Adding α,α -disubstituted and β -linked monomers to the genetic code of an organism

<https://doi.org/10.1038/s41586-023-06897-6>

Received: 29 April 2023

Accepted: 23 November 2023

Published online: 10 January 2024

 Open access

[Check for updates](#)

Daniel L. Dunkelmann^{1,2}, Carlos Piedrafita^{1,2}, Alexandre Dickson¹, Kim C. Liu¹, Thomas S. Elliott¹, Marc Fiedler¹, Dom Bellini¹, Andrew Zhou¹, Daniele Cervettini¹ & Jason W. Chin^{1✉}

The genetic code of living cells has been reprogrammed to enable the site-specific incorporation of hundreds of non-canonical amino acids into proteins, and the encoded synthesis of non-canonical polymers and macrocyclic peptides and depsipeptides^{1–3}. Current methods for engineering orthogonal aminoacyl-tRNA synthetases to acylate new monomers, as required for the expansion and reprogramming of the genetic code, rely on translational readouts and therefore require the monomers to be ribosomal substrates^{4–6}. Orthogonal synthetases cannot be evolved to acylate orthogonal tRNAs with non-canonical monomers (ncMs) that are poor ribosomal substrates, and ribosomes cannot be evolved to polymerize ncMs that cannot be acylated onto orthogonal tRNAs—this co-dependence creates an evolutionary deadlock that has essentially restricted the scope of translation in living cells to α -L-amino acids and closely related hydroxy acids. Here we break this deadlock by developing tRNA display, which enables direct, rapid and scalable selection for orthogonal synthetases that selectively acylate their cognate orthogonal tRNAs with ncMs in *Escherichia coli*, independent of whether the ncMs are ribosomal substrates. Using tRNA display, we directly select orthogonal synthetases that specifically acylate their cognate orthogonal tRNA with eight non-canonical amino acids and eight ncMs, including several β -amino acids, α,α -disubstituted-amino acids and β -hydroxy acids. We build on these advances to demonstrate the genetically encoded, site-specific cellular incorporation of β -amino acids and α,α -disubstituted amino acids into a protein, and thereby expand the chemical scope of the genetic code to new classes of monomers.

The genetic code of living cells has been reprogrammed to enable the site-specific incorporation of hundreds of non-canonical amino acids (ncAAs) into proteins^{1,7,8}, and the encoded synthesis of non-canonical polymers and macrocyclic peptides and depsipeptides^{2,3,9}. Despite remarkable progress, the monomers that can be site-specifically incorporated into proteins in cells have been essentially limited to α -L-amino acids with variant side chains, and closely related hydroxy acids. Although a wider range of monomers have been incorporated in *in vitro* translation reactions^{10–14} primarily into short peptides, these *in vitro* approaches cannot be extended to living cells. Under starvation conditions, the permissivity of an endogenous phenylalanyl-tRNA synthetase to a β -amino acid has been exploited for its low-level incorporation, in competition with phenylalanine, at all phenylalanine codons¹⁵; this approach generates a mixture of cellular proteins, is incompatible with quantitative, site-specific incorporation at a single position in response to a single codon, and is therefore fundamentally incompatible with reprogramming the genetic code.

The encoded, site-specific, incorporation of ncMs via cellular translation requires the creation of orthogonal aminoacyl-tRNA synthetase (aaRS)—orthogonal tRNA pairs. The orthogonal synthetase recognizes

a ncM but not the canonical amino acids present in the cell, and selectively transfers the ncM onto its cognate orthogonal tRNA, which is targeted to a blank codon (most commonly the amber stop codon). The ncM, once loaded onto the orthogonal tRNA, must also be a substrate for ribosomal polymerization (Extended Data Fig. 1). Current methods for engineering aaRSs that selectively acylate new monomers onto their cognate tRNAs rely on translational readouts^{4,6,16} and therefore require the monomers to be ribosomal substrates for incorporation, often at specific sites in proteins. Since many ncMs of interest are poor ribosomal substrates^{13,15,17–21}, this creates an evolutionary deadlock in cells; an orthogonal synthetase cannot be evolved to selectively acylate an orthogonal tRNA with ncMs that are poor ribosomal substrates, and ribosomes cannot be evolved to polymerize ncMs that cannot be selectively acylated onto orthogonal tRNAs. We realized that this deadlock might be broken by developing direct selections for orthogonal synthetases that selectively acylate their cognate orthogonal tRNAs with ncMs, independent of whether the ncMs are ribosomal substrates.

We previously described tRNA extension (tREX), a rapid method to determine the aminoacylation status of user-defined tRNAs from cells²².

¹Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. ²These authors contributed equally: Daniel L. Dunkelmann, Carlos Piedrafita. ✉e-mail: chin@mrc-lmb.cam.ac.uk

Here we develop derivatives of tREX that enable specific acylated tRNAs, isolated from cells, to be fluorescently labelled (fluoro-tREX) or captured (bio-tREX). We create strategies for producing split tRNA^{Pyl} (derived from *Methanosarcina mazei* pyrrolysl tRNA_{CUA}), which contains new 5' and 3' ends at the anticodon.

We connect the genotype responsible for acylation to the acylation itself by fusing the gene for *M. mazei* pyrrolysl-tRNA synthetase (hereafter referred to as PylRS) to a split tRNA^{Pyl} (stRNA^{Pyl}) gene, creating stmRNA^{Pyl} genes that encode stmRNA^{Pyl}, in which the 3' half of the stRNA^{Pyl} is fused to the PylRS mRNA. We demonstrate that we can selectively enrich—by more than 300-fold—stmRNAs encoding active PylRS variants with respect to attenuated activity variants, using bio-mREX (a variation of bio-tREX applied to the stmRNA).

Bio-mREX forms the basis of tRNA display, a method for discovering synthetases that acylate their cognate tRNAs with ncMs, independent of whether the ncMs are substrates for ribosomal translation. We use tRNA display to select orthogonal aaRSs that specifically acylate their cognate, orthogonal tRNAs with several β-amino acids, an α,α-disubstituted amino acid and a β-hydroxy acid. Moreover, we build on our advance to enable the site-specific co-translational incorporation of selected β-amino acids and α,α-disubstituted amino acids into a recombinant protein produced in *E. coli*; we produce milligrams of ncM-containing protein per litre of culture, and solve the structure of a β-amino acid-containing protein.

Detection and isolation of acylated tRNAs

We demonstrated that we could determine the aminoacylation status of a specific tRNA isolated from cells by periodate oxidation followed by selective, probe-mediated extension of non-oxidized tRNAs with nucleotide derivatives bearing a fluorophore (Cy5) or biotin (Fig. 1a).

We isolated tRNAs from cells expressing tRNA^{Pyl}_{CUA} in the presence and absence of PylRS and (N^ε-((tert-butoxy)carbonyl)-L-lysine (Bock, **1**)), a known and efficient substrate of PylRS. We oxidized the isolated tRNAs with sodium periodate and annealed a DNA probe, containing a 3' Cy3 and a 5' poly-G stretch, to the 3' end of tRNA^{Pyl}. We extended the free 3' end of tRNA^{Pyl}, resulting from aminoacylation-mediated protection from periodate oxidation, using Klenow fragment exo- and a dNTP mix in which dCTP was replaced with Cy5-labelled dCTP; the free 3' end of the tRNA acts as an RNA primer for Klenow fragment exo-, which uses the dNTPs to produce the reverse complement of the annealed DNA probe. We visualized the fluorescent signals following gel electrophoresis, specifically detecting a strong Cy5-labelled band corresponding to extended tRNA^{Pyl}_{CUA} from cells expressing PylRS and provided with Bock (**1**) (Fig. 1b). The Cy3 signal resulting from the tRNA–DNA probe hybrid provided a measure of tRNA abundance (Supplementary Fig. 2). These experiments demonstrated that our approach, using fluoro-tREX, enables the acylation of a tRNA to be followed via the generation of a fluorescent signal. We confirmed that fluoro-tREX has a wide dynamic range and can detect the activity of synthetase variants with low activities (Supplementary Fig. 3).

As we sought to use tRNA extension-based methods to detect tRNAs that were acylated with monomers beyond α-L-amino acids, we wanted to ensure that following periodate oxidation, a range of monomers could be hydrolysed from the tRNA—this hydrolysis is required to generate the acylation-dependent signal in tREX-based approaches. The rate of ester hydrolysis for deacylating an acylated tRNA is expected to be inversely related to the pK_a (the negative logarithm of the acid dissociation constant, K_a) of the carboxylic acid in the monomer that acylated the tRNA²³. We acylated tRNA^{Pyl}_{CUA} with the α-L-amino acid Bock (**1**), its hydroxy acid analogue, and its desamino carboxylic acid analogue²⁴ (Supplementary Fig. 4); the calculated pK_a of these monomers is about 2.3, 3.7, and 4.6 respectively²⁵. Although we detected acylation by the α-L-amino acid with our initial

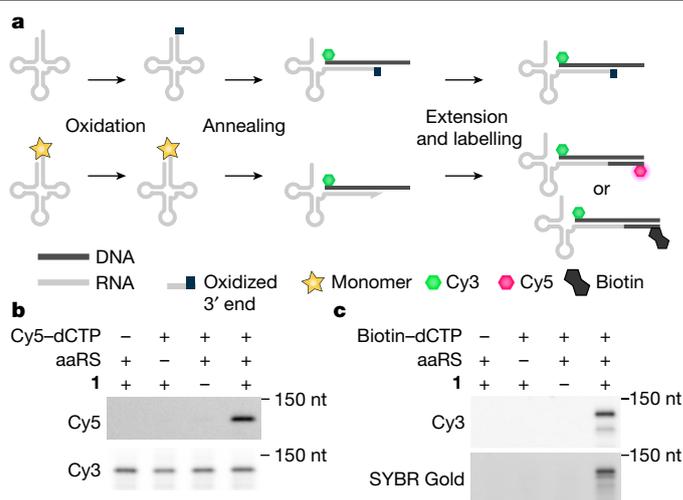


Fig. 1 | Acylation-dependent tRNA extension enables the sensitive detection and isolation of acylated tRNAs. **a**, Schematic of fluoro-tREX and bio-tREX protocols. tRNAs are isolated from cells, and the diol functionality of the 3' ribose on non-acylated tRNAs is oxidized to the dialdehyde. The acyl group of charged tRNAs protects the diol functionality of the 3' ribose and prevents oxidation to a dialdehyde. A Cy3-labelled DNA probe complementary to the 3' end of a target tRNA is annealed, and target tRNAs that were acylated are extended upon addition of Klenow fragment exo- and modified nucleotides. For fluoro-tREX, Cy5-labelled nucleotides are incorporated. Acylated tRNAs lead to a Cy5 and Cy3 signal, whereas non-acylated tRNAs only give a Cy3 signal. For bio-tREX, biotinylated nucleotides are incorporated and the tRNAs that were acylated are purified using streptavidin beads and can be visualized by SYBR gold staining following gel electrophoresis. **b**, Fluoro-tREX detected the acylation of tRNA^{Pyl}_{CUA} in the presence of PylRS and Bock (**1**). Cells expressing tRNA^{Pyl}_{CUA} were grown in the presence and absence of PylRS and Bock (**1**). **c**, Bio-tREX enables the selective isolation of previously acylated tRNAs. Cells harbouring tRNA^{Pyl}_{CUA} were grown in the presence and absence of PylRS and Bock (**1**). Isolation of the tRNA and associated probe was visualized by SYBR Gold staining for RNA and Cy3 fluorescence for the probe. Experiments in **b,c** were repeated three times with similar results. For full, uncropped gels for all figures see Supplementary Fig. 1.

fluoro-tREX protocol, we did not detect acylation of tRNAs by the hydroxy acid or simple carboxylic acid; we hypothesized that this difference in detection was due to incomplete deacylation of the tRNAs loaded with the hydroxy acid or simple carboxylic acid following oxidation, in our fluoro-tREX protocol. By treating tRNAs with base, following oxidation, we improved the detection of acylation for all monomers tested. These results suggested that our revised protocol would enable detection of tRNA acylation for an extended set of monomers.

Next, we replaced Cy5-dCTP with biotinylated dCTP (bio-dCTP) in the extension step of fluoro-tREX, thereby creating biotin-tREX (bio-tREX). We selectively captured the biotinylated extension product (resulting from tRNA molecules that were protected from periodate oxidation by their aminoacylation) on streptavidin beads, washed the beads, and eluted bound tRNA extension products. The presence of the tRNA^{Pyl}_{CUA} extension product in the eluate was dependent on the presence of PylRS and Bock (**1**) in cells, and the addition of bio-dCTP to the extension reaction (Fig. 1c). We conclude that bio-tREX enabled the selective capture of tRNA extension products from tRNAs that were acylated.

Production and acylation of split tRNAs

Next, we tested whether we could split the tRNA^{Pyl}_{CUA} gene at the anticodon to create a split tRNA^{Pyl} (stRNA^{Pyl}). We envisioned a system in which

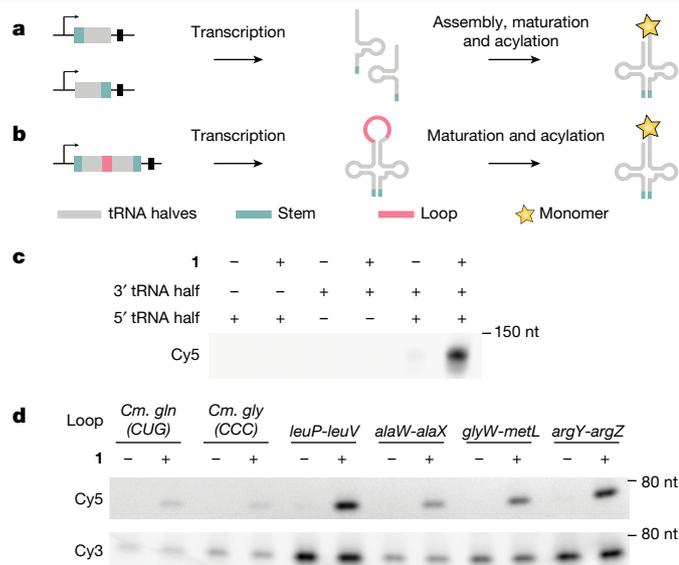


Fig. 2 | Production and acylation of split tRNAs expressed from split and circularly permuted genes. **a**, Schematic for producing split tRNAs *trans*. The tRNA gene is split at the anticodon loop and the anticodon stem sequence is extended for optimal assembly of the transcribed RNA *in vivo*; this creates two genes: one for the 5' half and one for the 3' half of the split tRNA. These genes are transcribed and the split tRNA is assembled, matured and acylated in cells. **b**, Schematic for producing split tRNAs *cis* from a single gene. The tRNA sequence is circularly permuted by connecting the 3' half, via an intervening loop sequence, to the 5' half, splitting the sequence at the anticodon and extending the anticodon stem. Transcription, assembly *in cis* and maturation leads to a functional split tRNA. **c**, *In vivo* transcription, assembly, maturation and acylation of split tRNA^{Pyl} produced from genes for the 5' half and 3' half. Cells were grown in the presence of PylRS. Only the expression of both tRNA^{Pyl} halves led to a Bock (1)-dependent acylation signal, as judged by fluoro-tREX. Note that under the purification conditions used to isolate these stRNAs, we do not observe the Cy3 probe. **d**, Circularly permuted split tRNA^{Pyl} with different loop sequences were assayed by fluoro-tREX. For the *argY-argZ* and *leuP-leuV* loops derived from the intergenic regions of pairs of tRNA genes in *E. coli*, the fluoro-tREX signal for split tRNA production (Cy3) and acylation (Cy5) was comparable to the corresponding signal for intact tRNA^{Pyl} (Supplementary Fig. 6). Experiments in **c,d** were repeated three times with similar results.

the 5' and 3' halves of a split tRNA gene were transcribed, assembled *in vivo* via non-covalent interactions (including base-pairing), matured by the cellular tRNA processing machinery, and recognized and efficiently acylated by PylRS, which does not recognize the anticodon^{26,27} (Fig. 2a,b).

We first designed a series of constructs in which we split the tRNA gene at the anticodon. We replaced the sequence of the anticodon stem loop in each half of the tRNA_{CUA}^{Pyl} gene with an extension of 0 to 14 nucleotides in length; the extensions within each pair of tRNA halves were designed to base pair with each other and form a stem to stabilize the stRNA^{Pyl}.

We expressed each pair of tRNA halves *trans* (from two different plasmids) in the presence or absence of PylRS and Bock (1). We analysed the acylation of stRNA^{Pyl} using fluoro-tREX, as a measure of stRNA^{Pyl} assembly and function. For the stRNA^{Pyl} molecules with a stem length of 0 and 8 base pairs (bp), we observed attenuated acylation, and these constructs may not stably assemble in cells (Supplementary Fig. 5). For stRNA^{Pyl} molecules with stems comprising more than 12 bp, we observed gel bands consistent with degradation products (Supplementary Fig. 5). Stems of 10 or 12 bp resulted in robust aminoacylation, which was dependent on the presence of both tRNA halves, PylRS and Bock (1) (Fig. 2c). We concluded that a 10-bp stem is sufficient to

facilitate the association of the 2 tRNA halves, minimize degradation and enable robust aminoacylation of the assembled tRNA. We therefore performed all subsequent experiments with stRNA^{Pyl} molecules with a 10-bp stem. To our knowledge, this is the first demonstration of a split tRNA being assembled, matured and acylated in cells.

Next, we focused on designing an expression system for stRNA^{Pyl} that would both ensure equimolar stoichiometry of both tRNA halves and facilitate spatial proximity, and thereby the assembly, of the two halves. We hypothesized that we could produce the two tRNA halves *cis* from one transcript by inserting an intervening sequence between them, thereby creating a circular permutation of the parent tRNA. Processing to remove the intervening sequence would yield the stRNA (Fig. 2b).

We noted that the intergenic regions of polycistronic tRNA operons in *E. coli* connect the 3' end of a tRNA gene to the 5' end of the following tRNA gene and are efficiently removed from the resulting transcripts²⁸. We used the tRNA operon generator²⁹ to select four *E. coli* intergenic regions (and also selected two intervening sequences from circularly permuted *Cyanidioschyzon merolae* tRNA genes³⁰) as the intervening sequences for our circular permutation strategy.

We created *cis* stRNA^{Pyl} genes in which the 3' half of the stRNA^{Pyl} gene was connected through each selected intervening 'loop' sequence to the 5' half of the split tRNA gene. Fluoro-tREX revealed efficient expression of stRNAs from *cis* stRNA^{Pyl} genes (Cy3 signal), and substantial, Bock (1)-dependent acylation of the stRNA^{Pyl} molecules produced from the *cis* stRNA^{Pyl} genes (Cy5 signal) in the presence of PylRS (Fig. 2d). *Cis* stRNA^{Pyl} genes with the intergenic regions from *E. coli* produced stRNA^{Pyl} at levels comparable to tRNA_{CUA}^{Pyl}, and stRNA^{Pyl} was acylated at comparable levels to tRNA_{CUA}^{Pyl} (Supplementary Fig. 6). We used the *E. coli* intergenic region *leuP-leuV* for all further experiments. We conclude that tRNA_{CUA}^{Pyl} can be split and expressed from a single transcript *cis*, and that the split tRNA functions as an efficient substrate for acylation by PylRS.

Linking acylation phenotype and genotype

Next, we fused the PylRS coding sequence and a linker sequence to the 5' end of the 3' half of the *cis* stRNA^{Pyl}. We put the resulting stmRNA^{Pyl} cassette (Fig. 3a) under the control of an inducible *T7* promoter to maximize its transcription. We hypothesized that transcription, processing and maturation of this construct would lead to an stmRNA in which the mRNA encoding the synthetase was covalently linked to the 5' end of the 3' half of the stRNA, and associated with the 5' half of the stRNA. Translation of the synthetase mRNA within the stmRNA^{Pyl} would generate the synthetase protein, which—in the presence of its cognate nCm—would acylate the stmRNA^{Pyl} in the same cell (Fig. 3a). This would generate a covalent link between the monomer attached to the tRNA and the mRNA of the synthetase gene that catalysed the attachment.

We grew cells expressing stmRNAs in the presence and absence of Bock (1), isolated total RNA and performed fluoro-mREX, a modification of fluoro-tREX for larger RNAs. For the wild-type stmRNA we observed a Bock (1)-dependent fluorescent (Cy5) band in fluoro-mREX (Fig. 3b). We did not observe a fluorescent signal for stmRNA^{at} (an stmRNA containing a PylRS gene encoding a protein with attenuated activity; Supplementary Fig. 7) in fluoro-mREX (Fig. 3b).

We conclude that our stmRNA construct is functional; it is transcribed and processed to generate a split tRNA in which the 3' half is fused to the synthetase mRNA. The synthetase mRNA is translated, and the resulting protein catalyses the acylation of the 3' end of the 3' half of the stmRNA. In fluoro-mREX the acylation is converted into a fluorescent 'phenotype', thereby creating a physical linkage between the mRNA sequence of the synthetase, its genotype, and the fluorescent phenotype, generated as a result of the activity of the synthetase.

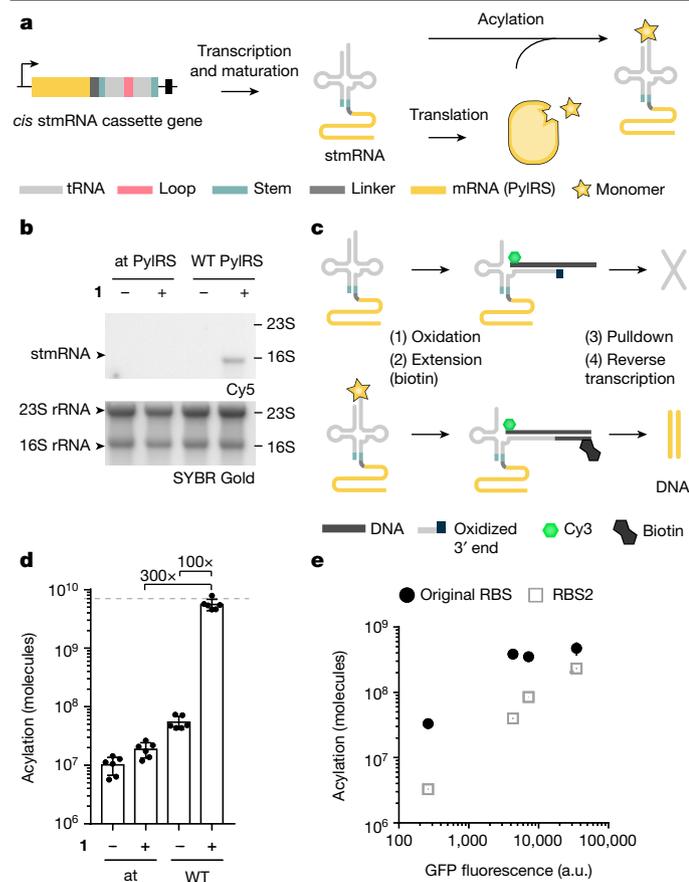


Fig. 3 | stmRNAs enable selective isolation of active PyIRS variants. **a**, Schematic of the *cis* split tRNA–mRNA fusion (stmRNA) gene, and the production and acylation of stmRNA. **b**, stmRNA acylation visualized by fluorescent mRNA extension (fluoro-mREX). Cells harboured an stmRNA gene, wild-type (WT) or attenuated (at) PyIRS. Positions of 16S (1.5 kb) and 23S rRNA (2.9 kb) are indicated. The fusion between the 3' half of the tRNA and the mRNA is 1.5 kb. The fluoro-mREX signal was visualized on denaturing gels; representative of three independent replicates. **c**, Schematic of biotin mRNA extension (bio-mREX). Biotinylated stmRNAs are enriched on streptavidin beads. The mRNA of PyIRS within is reverse transcribed on the beads and quantified by qPCR. **d**, Efficient and selective isolation of active PyIRS variant cDNA via bio-mREX. Cells harbouring the indicated stmRNA were grown in the presence or absence of Bock (1) and bio-mREX was performed. Following pulldown and reverse transcription, we determined the number of cDNA molecules. Dashed line indicates 2.5% of input (Supplementary Fig. 8). The bars represent the mean of five biological replicates, dots represent individual data points, and error bars show the s.d. **e**, Tailoring the RBS of PyIRS mRNAs within stmRNAs leads to a stronger correlation between the acylation of stmRNAs and readthrough of an amber stop codon (original RBS: $R^2 = 0.4694$, $P = 0.3148$; RBS2: $R^2 = 0.9742$, $P = 0.013$). Bio-mREX was performed from cells harbouring stmRNA genes encoding PyIRS(CbzK1–4) with either the original RBS or RBS2 grown in the presence of CbzK (2). The measured cDNA molecules were plotted against the fluorescence intensity of GFP(150CbzK)–His₆, resulting from readthrough of the amber codon in GFP(150TAG)His₆, by each PyIRS variant paired with tRNA^{Py}_{CUA} in cells provided with CbzK (2). Bio-mREX was performed in duplicates and GFP fluorescence was measured in triplicates. Error bars show the s.d. a.u., arbitrary units.

Acylation-specific enrichment of PyIRS

Next, we aimed to selectively isolate acylated stmRNA with respect to non-acylated stmRNA and reverse transcribe the isolated PyIRS mRNA within stmRNA to directly yield the cDNA of the PyIRS gene responsible for cellular acylation (Fig. 3c).

To selectively isolate acylated stmRNAs with respect to non-acylated stmRNAs we created bio-mREX (Fig. 3c), an adaptation of bio-tREX, in which we used the same methods to isolate RNA that we used for fluoro-mREX. In this approach, the biotinylated stmRNAs (which result from aminoacylated stmRNAs) are selectively captured on streptavidin beads. The PyIRS mRNAs are then directly reverse transcribed from the stmRNA captured on the beads, to create PyIRS cDNA. The PyIRS cDNA is then released by RNase H treatment and heating, and quantified by quantitative PCR (qPCR).

To test our approach, we grew *E. coli* cells harbouring the stmRNAs in the presence and absence of Bock (1) and performed bio-mREX. We recovered 100-fold more cDNA molecules for wild-type stmRNA in the presence of Bock (1) than in the absence of Bock (1) (Fig. 3d). Moreover, we recovered 300-fold more DNA molecules from wild-type stmRNA with Bock (1) than from stmRNA^{at} (Fig. 3d, Supplementary Fig. 8 and Supplementary Data 1). We concluded that bio-mREX enables the selective recovery of stmRNAs and the genes for synthetases that acylate the split tRNAs within them.

Additional experiments suggested that above a threshold activity, the first generation of bio-mREX could not effectively differentiate between PyIRS variants with different acylation activities (Fig. 3e, Extended Data Fig. 3, Supplementary Note 1, Supplementary Fig. 9 and Supplementary Data 1). To increase the dynamic range of the bio-mREX system, we attenuated the level of PyIRS protein by tuning the ribosome binding site (RBS) from which it is expressed within the stmRNA. We designed 5' untranslated region sequences with predicted attenuated translation rates³¹ and introduced them into the mRNAs for PyIRS variants of varying activity within stmRNA. For all new RBS sequences, the correlation between GFP expression resulting from amber suppression and the number of molecules recovered by bio-mREX increased with respect to the original construct (Fig. 3e, Supplementary Fig. 9 and Supplementary Data 1). RBS2 displayed a strong correlation and we proceeded to use the stmRNA utilizing this RBS, termed stmRNA^{vol2}, for all subsequent experiments.

In conclusion, by combining the stmRNA^{vol2} construct with bio-mREX, we developed a pulldown to selectively isolate the cDNA of active PyIRS variants and to effectively differentiate between variants with a range of activities.

tRNA display

We envisaged efficiently identifying PyIRS variants from large libraries of mutants (in stmRNA^{vol2}) that are active and selective for ncMs by running parallel bio-mREX-based selections, in the presence (positive sample) and absence (negative sample) of ncMs. The experiments would be performed in multiple replicates, and cDNA of the positive and negative samples, as well as the cDNA reverse transcribed from the input library, would subsequently be barcoded and sequenced by next-generation sequencing (NGS) (Fig. 4a). NGS data from the selections of the samples would enable the calculation of two key parameters: (1) the enrichment, defined as the average abundance of a particular sequence in the positive samples over the abundance of the same sequence in the input RNA; and (2) the selectivity, defined as the ratio of the abundance of a sequence in the positive sample over the abundance of the same sequence in the negative sample. Desired PyIRS variants would have sequences that are highly enriched and selective (Fig. 4a). We termed this selection approach tRNA display.

To test tRNA display we generated a small PyIRS library, in which we expected many sequences to be active, in stmRNA^{vol2}. The library targeted positions Y306, L309 and N346 in PyIRS, where mutations had previously been identified^{32,33} that enable the efficient incorporation of CbzK (2) (Fig. 4b and Extended Data Fig. 4). Following a single round of selection, we observed a large population of highly

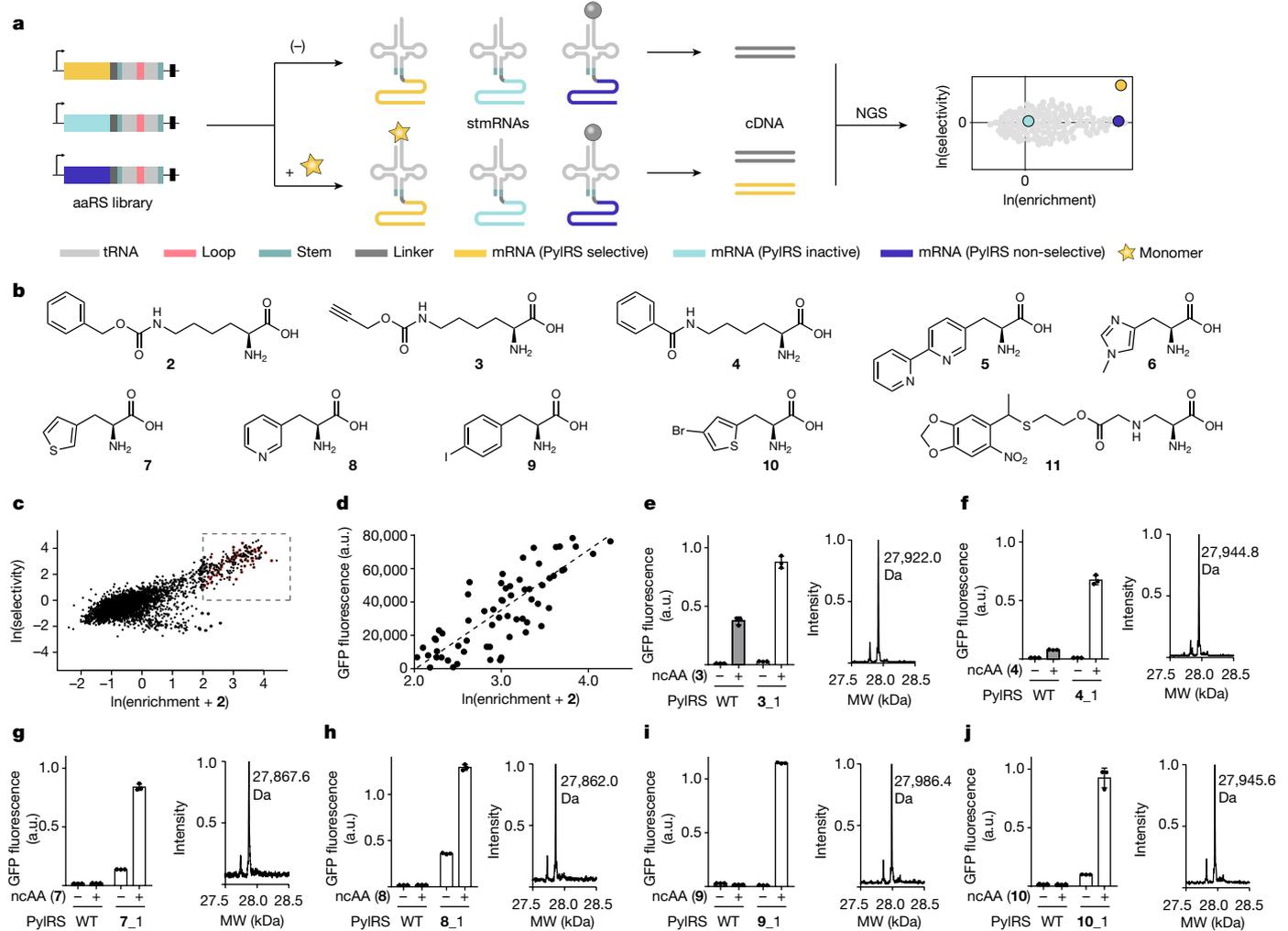


Fig. 4 | tRNA display enables the direct selection of orthogonal aminoacyl-tRNA synthetases that aminoacylate their cognate orthogonal tRNAs with ncAAs. **a**, Schematic of tRNA display. A library of aaRSs encoded within *stmRNA* genes is grown in the presence and absence of non-canonical monomers of interest (yellow star). Bio-mREX is performed and the cDNA is sequenced by NGS. The data are used to generate spindle plots. **b**, The numbered structures of non-canonical α -amino acids used in this study. **c**, tRNA display with *stmRNA*^{vol2_{lib1}}. Spindle plot from one step tRNA display selection with *stmRNA*^{vol2_{lib1}} and CbzK (2). Samples were run in triplicate and data were processed as described in Methods. Red dots indicate 65 clones that were further characterized. **d**, Plot of $\ln(\text{enrichment} + 2)$ of PyIRS mutants derived from tRNA display (red dots in **c**) against GFP fluorescence from cells containing the corresponding PyIRS mutant-tRNA_{CUA}^{PyI} pair, GFP(150TAG)His₆ and CbzK (2).

enriched and selective variant PyIRS sequences in the spindle plot derived from the sequencing of this tRNA display experiment (Fig. 4c).

To assess the predictive power of tRNA display, we measured the in vivo production of GFP(150CbzK)-His₆ (His₆-tagged GFP protein with a CbzK substitution at position 150) from GFP(150TAG)His₆ in the presence of tRNA_{CUA}^{PyI} and 65 PyIRS variants, which were hits on the basis of their position on the spindle plot (Supplementary Fig. 10). We observed a positive correlation between the enrichment derived from NGS data and the translation activity derived from GFP production for these hits (Fig. 4d), and found the vast majority of these hits to be selective (Supplementary Fig. 10). We concluded that tRNA display enables the direct, translation independent, identification of active and selective PyIRS enzymes from a library of PyIRS sequences.

The dotted line represents a linear regression for the data points. $R^2 = 0.6611$, $P < 0.0001$. **e-j**, Left, GFP fluorescence from cells containing GFP(150TAG)His₆, the indicated PyIRS variant-tRNA_{CUA}^{PyI} pair, and ncAA (white bar), or the wild-type PyIRS-tRNA_{CUA}^{PyI} pair with the same ncAA (grey bar). Fluorescence is plotted as a fraction of the signal generated by the wild-type PyIRS-tRNA_{CUA}^{PyI} pair with 2 mM Bock (1) and GFP(150TAG)His₆. Right, ESI-MS analysis of GFP(150X)-His₆, where X is the indicated ncAA. **e**, Found mass: 27,922.0 Da, expected mass 27,923.3 Da. **f**, Found mass: 27,944.8 Da, expected mass 27,945.5 Da. **g**, Found mass: 27,867.6 Da, expected mass 27,866.4 Da. **h**, Found mass: 27,862.0 Da, expected mass 27,861.4 Da. **i**, Found mass: 27,986.4 Da, expected mass 27,986.2 Da. **j**, Found mass: 27,945.6 Da, expected mass 27,944.3 Da. Bars represent the mean of three biological replicates, data points are shown as dots, and error bars represent s.d. Mass spectrometry data are from single replicates.

Scalable tRNA display-based discovery

To further validate the utility of tRNA display, we ran parallel selections (Supplementary Fig. 11) using six, highly diverse PyIRS active site libraries (Extended Data Fig. 4) and ten ncAAs (2-11) (Fig. 4b and Extended Data Fig. 2). After two rounds of selection, we analysed the spindle plots derived from the NGS data and identified individual PyIRS mutants for ncAAs 2, 3, 4, 7, 8, 9, 10 and 11 that were enriched and selective; the enriched and selective mutants for each of these ncAAs showed convergent sequence motifs (Supplementary Figs. 12-21).

We demonstrated the incorporation of ncAAs 3, 4, 7, 8, 9, 10 and 11 in response to the amber codon in GFP(150TAG)His₆ in cells containing tRNA_{CUA}^{PyI} and the corresponding PyIRS mutants identified by tRNA display. GFP production was dependent on the presence of ncAAs, and

electrospray ionization mass spectrometry (ESI-MS) confirmed the incorporation of each ncAA in GFP (Fig. 4e–j, Supplementary Figs. 12–21 and Supplementary Data 1). We note that our results include an aminoacyl-tRNA synthetase for **7**, which, to our knowledge, enables the first incorporation of this thiophene containing ncAA into a protein. Out of the 30 characterized variants with a selectivity score greater than or equal to 10 and an enrichment score greater than or equal to 5, 27 were active with their cognate ncAA in protein expression (20 variants were active at levels corresponding to at least 50% of the activity of PylRS with Bock (**1**)). All 27 variants selectively incorporated their ncAA substrate, as judged by mass spectrometry (Fig. 4e–j and Supplementary Figs. 12–21).

In summary, we demonstrated the parallel, scalable and rapid selection of PylRS variants using diverse libraries and ncAAs through tRNA display. We note that the amount of ncAA used in each tRNA display selection was 50–100 times lower than that used in current methods for synthetase selection.

Selective orthogonal pairs for ncMs

Next, we challenged tRNA display to discover synthetases that are selective for classes of ncMs that either cannot be translated or are likely to be poor ribosomal substrates^{13,15} (Fig. 5, Extended Data Fig. 2 and Supplementary Fig. 22).

To select PylRS variants for selected monomers, we ran a single round of parallel tRNA display selections (Supplementary Fig. 22) using a highly diverse library (lib14), which mutated residues 300, 302, 305, 306, 309, 344, 346, 348 and 401 in the PylRS active site, and a collection of ncMs (Extended Data Fig. 4). We analysed the spindle plots derived from the NGS data for selections with each ncM to identify PylRS variants that were enriched and selective (Supplementary Figs. 23–31). For ncMs **12** (6-((*tert*-butoxycarbonyl)amino)hexanoic acid (BocAhx)) and **13** ((*S*)-3-amino-3-(3-bromophenyl)propanoic acid (β^3 -mBrF)), we identified enriched and selective sequences, and found that the most selective hits converged on a distinct sequence pattern for each of these ncM.

The four most selective and active hits from the selection with **12**, PylRS(**12_1**) to PylRS(**12_4**), differ from the wild-type sequence by point mutations of V401. We showed, by fluoro-tREX, that all of these selected sequences were active and selective for **12** (Supplementary Fig. 23).

Notably, two PylRS variants, PylRS(**13_1**) and PylRS(**13_2**), identified by tRNA display selection, directed **13**-dependent acylation of tRNA_{CUA}^{Pyl} as judged by fluoro-tREX (Fig. 5c and Supplementary Fig. 24). To verify the identity of the monomer attached to the tRNA by PylRS(**13_1**) and PylRS(**13_2**), we captured the acylated tRNA_{CUA}^{Pyl} on streptavidin beads via a biotinylated probe for tRNA_{CUA}^{Pyl}, washed the beads and eluted the ncM by heating under alkaline conditions. We then derivatized the free ncM and analysed the sample by liquid chromatography–mass spectrometry (LC-MS) (Extended Data Fig. 5). Using this approach, we confirmed that both PylRS variants charged tRNA_{CUA}^{Pyl} with the ncM **13** (Fig. 5b and Supplementary Fig. 32).

Next we increased the activity of PylRS(**13_1**) and PylRS(**13_2**) by random mutagenesis of the active site region of the PylRS gene within the stmRNA construct followed by tRNA display-based selection (Supplementary Figs. 33 and 34). Three of the most enriched and selective hits from the resulting spindle plot, PylRS(**13_1**^{evol1–3}), were derivatives of PylRS(**13_1**). We confirmed the specificity of PylRS(**13_1**^{evol1–3}) for acylating tRNA_{CUA}^{Pyl} with **13**, by both fluoro-tREX and our LC-MS-based assay (Fig. 5b,c and Supplementary Figs. 32 and 34). PylRS(**13_1**^{evol1–3}) were notably more active in acylating tRNA_{CUA}^{Pyl} with **13** than PylRS(**13_1**) (Fig. 5b,c).

Next, we performed two rounds of selection with lib14 and ncMs **14**, **15**, **16**, **17** and **18** (β -amino acids with different side chains), **19** (an α,α -disubstituted amino acid) and **20** (a β -hydroxy acid) (Fig. 5a and Supplementary Fig. 35). From the resulting spindle plots

(Supplementary Figs. 36–42), we identified enriched and selective sequences, and the most selective hits converged on a distinct sequence pattern for each ncM. We note that the synthetases selected for all six β -amino acids differ in sequence, but contain common mutations M300D and A302H (Supplementary Figs. 25 and 38–42). The sequence pattern observed for the β -hydroxy acid **20** is similar to the one observed for β -amino acids. However, the residue at position 300—which may be in direct proximity to the amine or hydroxy group—is changed from aspartic acid to asparagine (Supplementary Fig. 42). The PylRS variants identified by tRNA display selection directed the specific acylation of tRNA_{CUA}^{Pyl} by their cognate monomer, as judged by fluoro-tREX and our LC-MS-based assay (Fig. 5d–q). We quantified the fraction of acylation as a function of ncM concentration (Supplementary Fig. 43 and Supplementary Data 1). To our knowledge, this is the first report of specific aminoacyl-tRNA synthetase–tRNA pairs for three distinct classes of ncM: β -amino acids, α,α -disubstituted amino acids and β -hydroxy acids.

Encoding ncMs in proteins

To investigate the incorporation of the β -amino acids, α,α -disubstituted amino acids and β -hydroxy acids into proteins we added the orthogonal synthetase–orthogonal tRNA pairs for **13**–**20**, the cognate ncMs, and GFP(150TAG)His₆ to cells. We observed ncM-dependent GFP production for **13**, **15**, **18** and **19**, with isolated yields ranging from 3 to 35 mg per litre of culture, and mass spectrometry confirmed the incorporation of these β -amino acids and α,α -disubstituted amino acids in GFP (Fig. 5r–t and Supplementary Figs. 44 and 45). In the absence of **13**, we observed some GFP production resulting from the incorporation of natural amino acids (Fig. 5r). However, in the presence of **13**, we produced more GFP, and we only detected incorporation of **13** by intact MS and MS/MS (Fig. 5s and Supplementary Fig. 44). We concluded that in the presence of **13**, the background incorporation observed in the absence of **13** was effectively outcompeted. We made similar observations for **15**, **18** and **19** (Fig. 5s,t and Supplementary Fig. 44). Similar observations have previously been made for efficient and selective ncAA incorporation systems³⁴ and the fidelity of the natural code also relies on competition³⁵. We concluded that **13**, **15**, **18** and **19** are site-specifically incorporated with high fidelity. We did not observe incorporation of ncMs **13**, **15**, **18** or **19** at position 3 of GFP (from GFP(3TAG)His₆) (Supplementary Fig. 45), indicating that these ncMs are not tolerated at all positions in a protein. Similar site-dependent incorporation efficiency has previously been observed for ncAAs^{36,37}.

We note that we did not observe an ncM-dependent increase in production of GFP from GFP(150TAG)His₆ or GFP(3TAG)His₆ with **14**, **16**, **17** or **20** when cells were provided with these ncMs and their cognate orthogonal synthetase–orthogonal tRNA pairs (Fig. 5r and Supplementary Fig. 45). These observations are consistent with these ncMs being poor substrates for ribosomal polymerization. For **17** and **20**, we observed a decrease in GFP production upon addition of ncM (Fig. 5r). This is consistent with these ncMs, once acylated onto the orthogonal tRNA, inhibiting readthrough of the amber codon. The discovery of orthogonal synthetases that are specific for these ncMs provides a starting point for selecting ribosomes that can efficiently polymerize them.

Structure of β -amino acid-containing protein

To further characterize the incorporation of **13** at position 150 of GFP, we solved the structure of GFP(150(S) β^3 mBrF)–His₆ at 1.5 Å resolution by X-ray crystallography (the protein was purified from cells harbouring PylRS(**13_1**) and tRNA_{CUA}^{Pyl}) (Fig. 5u and Extended Data Table 1). The electron density shows consecutive carbon atoms (C2 and C3) in the protein backbone at position 150; the meta-bromophenyl substituent is attached to C3 of the β -amino acid, and the stereochemistry at C3 corresponds to the expected (S) stereoisomer. Our structure confirms the site-specific incorporation of the expected (S) β^3 mBrF-monomer in

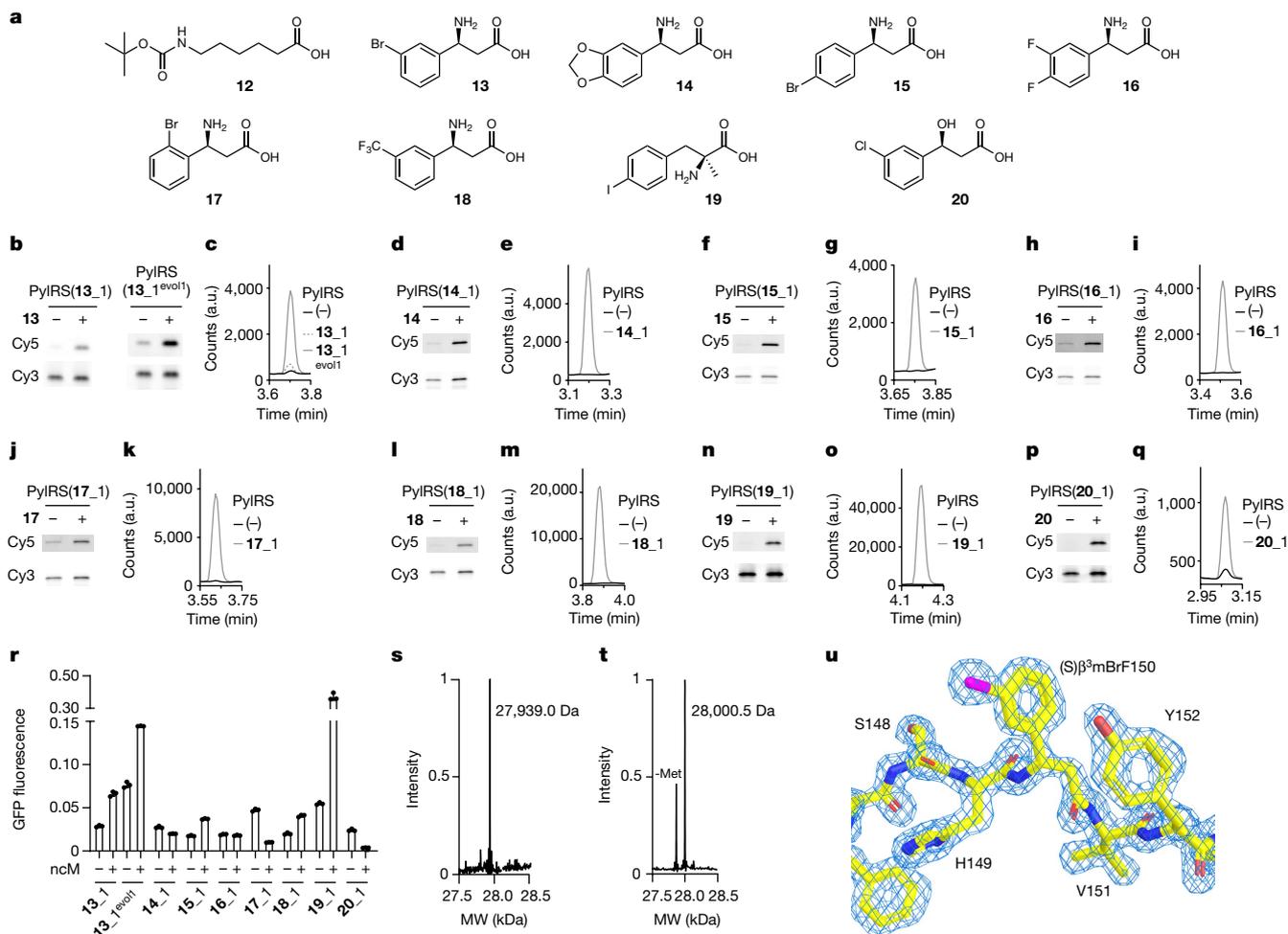


Fig. 5 | tRNA display selection of orthogonal synthetases that charge ncMs.

a, ncMs for which selective PyIRS mutants were discovered. **b**, Fluoro-tREX. A representative gel is shown for each PyIRS variant. Experiments performed in independent triplicates with similar results. **c**, Selected PyIRS variants acylate tRNA^{Pyl}_{CUA} with **13**. LC-MS traces (scanning ion mode on 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC) adduct of **13**) of AQC-derivative eluted from tRNA pull-downs. Cells harbouring the corresponding PyIRS variant and tRNA^{Pyl}_{CUA} (or only tRNA^{Pyl}_{CUA} (-)) were grown with **13**. Pull-downs used a biotinylated probe against tRNA^{Pyl}_{CUA}. Representative traces are shown. **d, f, h, j, l, n, p**. As in **b**, but with indicated PyIRS variants and **14** (**d**), **15** (**f**), **16** (**h**), **17** (**j**), **18** (**l**), **19** (**n**) or **20** (**p**). **e, g, i, k, m, o, q**. As in **c**, but with indicated PyIRS variants and **14** (**e**), **15** (**g**), **16** (**i**), **17** (**k**), **18** (**m**), **19** (**o**) or **20** (**q**). **e, g, i**. Performed in duplicate, all replicates yielded similar results. **c, k, m, o, q**. Performed in triplicate, all replicates yielded similar results. **r**, Fluorescence from cells containing

GFP(150TAG)His₆, indicated PyIRS variant and tRNA^{Pyl}_{CUA}, and grown in the presence or absence of ncM (**13–20**, 4 mM). Fluorescence is shown as a fraction of the fluorescence generated by the wild-type PyIRS–tRNA^{Pyl}_{CUA} pair with 4 mM Bock (**1**) and GFP(150TAG)His₆. Bar graphs represent mean of three independent measurements, individual data points are shown as dots and error bars indicate s.d. **s**, ESI-MS of GFP(150(S)^{β3}mBrF)–His₆ purified from cells harbouring PyIRS(**13**_{evol1}), tRNA^{Pyl}_{CUA} and GFP(150TAG)His₆ grown with **13**. Found mass: 27,939.0 Da, predicted mass: 27,938.2 Da. Spectra acquired once. **t**, ESI-MS of GFP(150(S)α-Me-pIF)–His₆ purified from cells harbouring PyIRS(**19**₁), tRNA^{Pyl}_{CUA} and GFP(150TAG)His₆ grown with **19**. Found mass: 28,000.5 Da, predicted mass: 28,000.2 Da. Spectra were acquired once. **u**, Close up on residue 150 of GFP(150(S)^{β3}mBrF)–His₆, from a crystal structure determined at 1.5 Å. The 2F_o – F_c map is shown at contour level of σ = 2 (Protein Data Bank (PDB) ID 8OVY), electron density (blue).

the protein. The introduction of (S)^{β3}mBrF leads to a notable kink in the beta strand of GFP, when compared to the wild-type protein (Extended Data Fig. 6). Notably, the hydrogen bonding networks of the residues immediately preceding and following the β-amino acid in the polypeptide chain remain essentially unperturbed; this indicates that this beta strand can accommodate the β³-amino acid at this position (Extended Data Fig. 6). To our knowledge, this represents the first solved structure of a protein produced in vivo that contains a genetically encoded β-amino acid.

Discussion

The in vivo, site-specific incorporation of backbone-modified monomers is a longstanding challenge in expanding the scope of encoded cellular polymer synthesis beyond α-L-amino acids and their close analogues².

tRNA display systematically breaks the deadlock (Extended Data Fig. 1) that has so far limited the range of monomers that can be used to specifically acylate orthogonal tRNAs in vivo. Using tRNA display, we have identified orthogonal synthetase–orthogonal tRNA pairs that are selective for eight new ncMs, including β-amino acids, α,α-disubstituted amino acids and β-hydroxy acids, and thereby directly facilitated the genetic encoding and site-specific incorporation of β-amino acids and α,α-disubstituted amino acids into proteins in a living organism.

tRNA display may be extended to other synthetase and tRNA systems to further expand the range of monomers that can be loaded onto tRNAs. The sequence, activity and selectivity data generated from tRNA display may facilitate the de novo design of active sites that are selective for new ncMs. Extensions of tRNA display may be used to select for genes that direct the biosynthesis of ncMs, or that bind to and protect tRNAs acylated with ncMs (for example, EF-Tu variants for ncMs^{38–40}).

In future work we will leverage the cellular acylation of tRNAs with ncMs to enable translation-based selections for orthogonal ribosomes^{34,41–43} that can polymerize ncMs that are poor substrates for natural translation, and facilitate the encoded cellular synthesis of polymers composed of more diverse ncMs. The repertoire of non-canonical polymers may be further enhanced by leveraging post-translational modifications and ligations^{44–49}.

The genetic encoding of β -amino acids and α,α -disubstituted amino acids may enable the creation of protease-resistant proteins and new drug-like molecules^{50,51} in living cells. Moreover, it may be possible to encode the biosynthesis of foldamers⁵² entirely composed of β -amino acids and other ncMs⁵³ that complement and augment the canonical functions of living organisms.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-023-06897-6>.

- Dumas, A., Lercher, L., Spicer, C. D. & Davis, B. G. Designing logical codon reassignment—expanding the chemistry in biology. *Chem. Sci.* **6**, 50–69 (2015).
- Robertson, W. E. et al. Sense codon reassignment enables viral resistance and encoded polymer synthesis. *Science* **372**, 1057–1062 (2021).
- Spinck, M. et al. Genetically programmed cell-based synthesis of non-natural peptide and depsipeptide macrocycles. *Nat. Chem.* **15**, 61–69 (2023).
- Santorio, S. W., Wang, L., Herberich, B., King, D. S. & Schultz, P. G. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nat. Biotechnol.* **20**, 1044–1048 (2002).
- Chin, J. W. et al. An expanded eukaryotic genetic code. *Science* **301**, 964–967 (2003).
- Chin, J. W. Expanding and reprogramming the genetic code of cells and animals. *Annu. Rev. Biochem.* **83**, 379–408 (2014).
- Young, D. D. & Schultz, P. G. Playing with the molecules of life. *ACS Chem. Biol.* **13**, 854–870 (2018).
- Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **550**, 53–60 (2017).
- De La Torre, D. & Chin, J. W. Reprogramming the genetic code. *Nat. Rev. Genet.* **22**, 169–184 (2021).
- Ellman, J. A., Mendel, D. & Schultz, P. G. Site-specific incorporation of novel backbone structures into proteins. *Science* **255**, 197–200 (1992).
- Mendel, D., Cornish, V. W. & Schultz, P. G. Site-directed mutagenesis with an expanded genetic code. *Annu. Rev. Biophys. Biomol. Struct.* **24**, 435–462 (1995).
- Hecht, S. M. Expansion of the genetic code through the use of modified bacterial ribosomes. *J. Mol. Biol.* **434**, 167211 (2022).
- Katoh, T. & Suga, H. In vitro genetic code reprogramming for the expansion of usable noncanonical amino acids. *Annu. Rev. Biochem.* **91**, 221–243 (2022).
- Fricke, R. et al. Expanding the substrate scope of pyrrolysyl-transfer RNA synthetase enzymes to include non- α -amino acids in vitro and in vivo. *Nat. Chem.* **15**, 960–971 (2023).
- Melo Czekster, C., Robertson, W. E., Walker, A. S., Soll, D. & Schepartz, A. In vivo biosynthesis of a β -amino acid-containing protein. *J. Am. Chem. Soc.* **138**, 5194–5197 (2016).
- Chin, J. W., Martin, A. B., King, D. S., Wang, L. & Schultz, P. G. Addition of a photocrosslinking amino acid to the genetic code of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 11020–11024 (2002).
- Tan, Z., Forster, A. C., Blacklock, S. C. & Cornish, V. W. Amino acid backbone specificity of the *Escherichia coli* translation machinery. *J. Am. Chem. Soc.* **126**, 12752–12753 (2004).
- Pavlov, M. Y. et al. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc. Natl Acad. Sci. USA* **106**, 50–54 (2009).
- Katoh, T., Tajima, K. & Suga, H. Consecutive elongation of D-amino acids in translation. *Cell Chem. Biol.* **24**, 46–54 (2017).
- Katoh, T. & Suga, H. Ribosomal incorporation of consecutive β -amino acids. *J. Am. Chem. Soc.* **140**, 12159–12167 (2018).
- Dedkova, L. M. et al. β -Puromycin selection of modified ribosomes for in vitro incorporation of beta-amino acids. *Biochemistry* **51**, 401–415 (2012).
- Cervellini, D. et al. Rapid discovery and evolution of orthogonal aminoacyl-tRNA synthetase-tRNA pairs. *Nat. Biotechnol.* **38**, 989–999 (2020).
- Barton, P., Laws, A. P. & Page, M. I. Structure-activity relationships in the esterase-catalysed hydrolysis and transesterification of esters and lactones. *Perkin Trans. 2*, 2021–2029 (1994).
- Kobayashi, T., Yanagisawa, T., Sakamoto, K. & Yokoyama, S. Recognition of non- α -amino substrates by pyrrolysyl-tRNA synthetase. *J. Mol. Biol.* **385**, 1352–1360 (2009).
- Pan, X., Wang, H., Li, C., Zhang, J. Z. & Ji, C. MolGpka: a web server for small molecule p_K prediction using a graph-convolutional neural network. *J. Chem. Inf. Model.* **61**, 3159–3165 (2021).
- Ambrogelly, A. et al. Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. *Proc. Natl Acad. Sci. USA* **104**, 3141–3146 (2007).
- Elliott, T. S. et al. Proteome labeling and protein identification in specific tissues and at specific developmental stages in an animal. *Nat. Biotechnol.* **32**, 465–472 (2014).
- El Yacoubi, B., Bailly, M. & de Crécy-Lagard, V. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu. Rev. Genet.* **46**, 69–95 (2012).
- Dunkelmann, D. L., Oehm, S. B., Beattie, A. T. & Chin, J. W. A 68-codon genetic code to incorporate four distinct non-canonical amino acids enabled by automated orthogonal mRNA design. *Nat. Chem.* **13**, 1110–1117 (2021).
- Soma, A. et al. Permuted tRNA genes expressed via a circular RNA intermediate in *Cyanidioschyzon merolae*. *Science* **318**, 450–453 (2007).
- Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).
- Yanagisawa, T. et al. Structural basis for genetic-code expansion with bulky lysine derivatives by an engineered pyrrolysyl-tRNA synthetase. *Cell Chem. Biol.* **26**, 936–949. e913 (2019).
- Dunkelmann, D. L., Willis, J. C. W., Beattie, A. T. & Chin, J. W. Engineered triply orthogonal pyrrolysyl-tRNA synthetase/tRNA pairs enable the genetic encoding of three distinct non-canonical amino acids. *Nat. Chem.* **12**, 535–544 (2020).
- Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.* **25**, 770–777 (2007).
- Fredens, J. et al. Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569**, 514–518 (2019).
- Chemla, Y., Ozer, E., Algov, I. & Alfanta, L. Context effects of genetic code expansion by stop codon suppression. *Curr. Opin. Chem. Biol.* **46**, 146–155 (2018).
- Pott, M., Schmidt, M. J. & Summerer, D. Evolved sequence contexts for highly efficient amber suppression with noncanonical amino acids. *ACS Chem. Biol.* **9**, 2815–2822 (2014).
- Doi, Y., Ohtsuki, T., Shimizu, Y., Ueda, T. & Sisido, M. Elongation factor Tu mutants expand amino acid tolerance of protein biosynthesis system. *J. Am. Chem. Soc.* **129**, 14458–14462 (2007).
- Ohtsuki, T., Yamamoto, H., Doi, Y. & Sisido, M. Use of EF-Tu mutants for determining and improving aminoacylation efficiency and for purifying aminoacyl tRNAs with non-natural amino acids. *J. Biochem.* **148**, 239–246 (2010).
- Achenbach, J. et al. Outwitting EF-Tu and the ribosome: translation with D-amino acids. *Nucleic Acids Res.* **43**, 5687–5698 (2015).
- Rackham, O. & Chin, J. W. A network of orthogonal ribosome \times mRNA pairs. *Nat. Chem. Biol.* **1**, 159–166 (2005).
- Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441–444 (2010).
- Schmied, W. H. et al. Controlling orthogonal ribosome subunit interactions enables evolution of new function. *Nature* **564**, 444–448 (2018).
- Lakis, E., Magyari, S. & Piel, J. In vivo production of diverse β -amino acid-containing proteins. *Angew. Chem.* **134**, e202202695 (2022).
- Camarero, J. A. & Muir, T. W. Native chemical ligation of polypeptides. *Curr. Protoc. Protein Sci.* **15**, 18.14. 11–18.14. 21 (1999).
- Niquille, D. L. et al. Nonribosomal biosynthesis of backbone-modified peptides. *Nat. Chem.* **10**, 282–287 (2018).
- Arnison, P. G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
- Wright, T. H. et al. Posttranslational mutagenesis: a chemical strategy for exploring protein side-chain diversity. *Science* **354**, aag1465 (2016).
- Roe, L. et al. Backbone extension acyl rearrangements enable cellular synthesis of proteins with internal β -peptide linkages. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.03.560714> (2023).
- Hintermann, T. & Seebach, D. The biological stability of β -peptides: no interactions between α - and β -peptidic structures? *Chimia* **51**, 244–244 (1997).
- Köksch, B., Sewald, N., Hofmann, H. J., Burger, K. & Jakubke, H. D. Proteolytically stable peptides by incorporation of α -Tfm amino acids. *J. Peptide Sci.* **3**, 157–167 (1997).
- Gellman, S. H. Foldamers: a manifesto. *Acc. Chem. Res.* **31**, 173–180 (1998).
- Wang, P. S. & Schepartz, A. β -Peptide bundles: Design. Build. Analyze. *Biosynthesis* **52**, 7420–7432 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

Methods

Buffers

Resuspension buffer: 100 mM sodium acetate, 50 mM NaCl, 0.1 mM EDTA, pH 5.0; deacylation buffer: 50 mM bicine pH 9.6, 1 mM EDTA; Buffer D: 50 mM sodium acetate pH 5, 150 mM NaCl, 10 mM MgCl₂, 0.1 mM EDTA; hybridization buffer: 10 mM Tris-HCl, 25 mM NaCl, pH 7.4; Klenow fragment exo- master mix with Cy5-11-dCTP (KMM-Cy5): 17 µl water, 5 µl 10x NEBuffer 2.0, 1 µl Klenow fragment exo- (NEB), 1 dNTPS-dCTP (10 mM), 1 µl Cy5-11-dCTP (20 µM); orange loading dye: 8 M urea, orange G; Klenow fragment exo- master mix mini (KMM-mini): 3.35 µl water, 1.2 µl 10x NEBuffer 2.0, 0.1 µl Klenow fragment exo-, 0.25 dNTPS (10 mM); Klenow fragment exo- master mix with Cy5-11-dCTP mini (KMM-Cy5-mini): 2.1 µl water, 1.2 µl 10x NEBuffer 2.0, 0.2 µl Klenow fragment exo-, 0.5 dNTPS (5 mM, without dCTP), 1 µl Cy5-11-dCTP (5 µM); Klenow fragment exo- master mix with Bio-11-dCTP (KMM-Bio): 17 µl water, 5 µl 10x NEBuffer 2.0, 1 µl Klenow fragment exo-, 1 dNTPS-dCTP (10 mM), 1 µl Bio-11-dCTP (20 µM); washing buffer: 10 mM Tris-HCl, 150 mM LiCl, 1 mM EDTA, 0.05% (v/v) Tween-20, pH 7.5; binding buffer: 20 mM Tris-HCl, 1 M LiCl, 2 mM EDTA, pH 7.5; formamide loading buffer (FLB): 90% formamide; SDS lysis buffer (SLB): 100 mM sodium acetate, 50 mM NaCl, 0.1 mM EDTA, pH 5.0, 1% (v/v) sodium dodecyl sulphate (SDS); alkaline washing buffer (AWB): 25 mM NaOH, 4 mM EDTA, 0.05% (v/v) Tween-20; reverse transcription hybridization mix (RHM): 1 µl DNA primer (2 µM), 1 µl 10 mM dNTPs, 1 µl 10x hybridization buffer, 10 µl water; reverse transcription master mix (RMM): 4 µl SSIV buffer, 1 µl RNase Out, 1 µl SSIV reverse transcriptase, 1 µl 0.1 M dithiothreitol (DTT); acidic washing buffer 1 (aWB1): 100 mM sodium acetate pH 5; acidic washing buffer 1 plus Tween-20 (aWB1-T): 100 mM sodium acetate pH 5, 0.01% (v/v) Tween-20; acidic washing buffer 2 (aWB2): 20 mM sodium acetate pH 5.

Media

SOC: Super Optimal Broth plus 20 mM glucose; 2xYT-s: Yeast Extract Tryptone supplemented with 75 µg ml⁻¹ spectinomycin; 2xYT-s-t: Yeast Extract Tryptone supplemented with 75 µg ml⁻¹ spectinomycin and 10 µg ml⁻¹ tetracycline; 2xYT-s-ap: Yeast Extract Tryptone supplemented with 75 µg ml⁻¹ spectinomycin and 50 µg ml⁻¹ apramycin; 2xYT-am: Yeast Extract Tryptone supplemented with 75 µg ml⁻¹ ampicillin.

Chemicals

NcAA **1** and **2** were purchased from Bachem. NcAA **4** was purchased from Fluorochem. NcAA **5** was purchased from Ambeed. NcAAs and ncMs **6**, **8**, **14**, **S1**, **S4** and **S5** were purchased from Amamine. NcAA **7** was purchased from aaBlocks. NcAAs and ncMs **9**, **10**, **12** and **S3** were purchased from Merck. NcMs **13**, **15**, **16**, **17**, **18**, **20** and **S7** were purchased from BLD. NcM **S2** was purchased from Advanced ChemBlock. NcM **19** and **S6** were purchased from AstaTech. NcAAs and ncMs **3** and **S8** were synthesized as previously described³, and ncAA **11** was custom synthesized as previously described³⁴. NcMs **S1** and **S5** were Boc-deprotected in concentrated HCl in dioxane.

DNA oligonucleotides

See Supplementary Data 2.

DNA constructs

See Supplementary Data 2 for all constructs.

DNA construct cloning

Standard cloning was performed by Gibson assembly using NEBuilder HiFi DNA Assembly Master Mix (NEB) according to manufacturer's guidelines. Libraries were generated by enzymatic inverse PCR, as previously described^{55,56}. In brief, a template plasmid was amplified by PCR using two primers (see primer list) containing degenerate codons at

desired mutagenesis sites and a BsaI cleavage site. In the case of custom mixes, primers containing different codons were manually mixed and used for PCR reactions. PCR products were gel purified and digested using BsaI and DpnI. Subsequently, samples were purified, ligated using T4 DNA Ligase, and transformed into electrocompetent *E. coli* DH10β cells ensuring a minimal transformation efficiency of 10⁹. Individual colonies (>10) were evaluated using Sanger sequencing for quality control of the library assembly. Total plasmid DNA was prepared from the resulting culture, sequenced as a bulk using Sanger sequencing and used for subsequent experiments.

General protocols

Isolation and oxidation of tRNAs (protocol A). This protocol was used to isolate tRNAs from 1–10 ml of cell culture. Chemically competent DH10β cells were transformed with a pMB1 plasmid encoding a PylRS variant and a tRNA, or a circularly permuted split tRNA, and rescued in 1 ml of SOC for 1 h at 37 °C, 700–1,000 rpm. Cells were transferred into selective 2xYT-s medium and grown overnight. Overnight cultures were diluted in a ratio between 1:20 and 1:40 and grown to OD₆₀₀ of 0.5–1. Cells were centrifuged at 4,200 rcf at 4 °C for 12 min, taken up in 200 µl resuspension buffer, transferred to a 96-well plate and centrifuged at 4,200 rcf at 4 °C for 12 min. Cells were resuspended in 135 µl resuspension buffer and 15 µl liquid phenol was added. Cells were lysed by shaking at 650 rpm for 20 min, and then centrifuged at 4,200 rcf at 4 °C for 20 min; the cell lysate was added to 40 µl chloroform, and the resulting suspension mixed by pipetting up and down. The mixture was centrifuged at 4,200 rcf at 4 °C for 10 min, and 115 µl of the aqueous layer were transferred into 6 µl 0.1 M NaIO₄. The isolated RNA was oxidized for 1 h on ice, and the oxidation reaction was quenched by addition of 8 µl of 0.1 M DTT. tRNAs were purified using the ZR-96 Oligo Clean & Concentrator from Zymo Research. In brief, 250 µl oligo binding buffer was added to the oxidation reaction, subsequently 400 µl isopropanol was added, and the mixture transferred to a 96-well silica column plate. The plate was centrifuged for 2 min, 4,200 rcf at room temperature, and 800 µl of oligo wash buffer was added. The plate was centrifuged for 2 min, 4,200 rcf at room temperature, aerated, and centrifuged for another 4 min, 4,200 rcf at room temperature. Finally, the RNA was eluted in either 14 µl water, when the samples were not processed, or in 50 µl water, when the samples were further deacylated, by centrifugation for 4 min, 4,200 rcf at room temperature.

Isolation and oxidation of tRNAs (protocol B). The volumes described in this protocol were used to isolate tRNAs from 5–25 ml of cell culture as previously described^{22,57}. In brief, cells were grown as described in protocol A, washed with 800 µl resuspension buffer and transferred to a 1.5 ml Eppendorf tube. Cells were taken up in 225 µl resuspension buffer and 25 µl of liquid phenol was added. Cells were lysed by vortexing for one minute and incubation, with head-over-tail rotation, for 20 min. Lysed cells were centrifuged for 15 min, 20,000 rcf at room temperature, the cell lysate was added to 250 µl chloroform, the samples were vortexed for one minute and then centrifuged, 10 min, 20,000 rcf at room temperature. Two-hundred microlitres of the aqueous layer was transferred into 10 µl 0.1 M NaIO₄ and the RNA was oxidized for 1 h on ice. Finally, the oxidized RNA was added to 440 µl ethanol and precipitated for at least 20 min at –20 °C. The samples were centrifuged for 25 min, 20,000 rcf at 4 °C and aspirated. RNA pellets were dried for 10 min at room temperature and dissolved in water or buffer D.

tRNA deacylation. 45 µl of isolated RNA was added to 5 µl 10 x deacylation buffer and tRNAs were deacylated for 36 min at 42 °C. The deacylation reaction was quenched by addition of 6 µl 3 M sodium acetate, and tRNAs were purified using the ZR-96 Oligo Clean & Concentrator from Zymo Research, as described in protocol A for the isolation and oxidation of tRNAs (with the exception of using 100 µl of oligo binding buffer, instead of 250 µl). Deacylated tRNAs were eluted in 14 µl water.

Article

Fluoro-tREX. This protocol was used to run the experiments shown in Fig. 2c and Supplementary Fig. 4. RNA concentrations were adjusted to the lowest common denominator and 10 μl of RNA was added to 2.5 μl 10x hybridization buffer, 11.5 μl water and 1 μl Cy3-labelled extension primer (2 μM ; note that probes for tREX-based approaches are described as primers throughout the methods even though they template extension and are not themselves extended). The DNA primer was hybridized at 65 °C for 5 min, before addition of 25 μl KMM-Cy5 and extension at 37 °C for 6 min. Samples were purified using the 10 μg NEB Monarch RNA clean-up Kit (NEB) and eluted in 12 μl water. 12 μl of orange loading dye was added, the samples were loaded onto a Novex TBE 6 M urea 10 or 15% PAGE gel (Invitrogen) and run for 36 min in 0.5x Tris-borate-EDTA (TBE) buffer at 270 V. Gels were imaged on an Amersham Typhoon Biomolecular Imager (GE) using the Cy3 and Cy5 emission filters. Then gels were stained with SYBR Gold (Invitrogen) and imaged again using the same filters.

Mini-fluoro-tREX. Unless stated otherwise, all fluoro-tREX experiments were run with the mini-fluoro-tREX protocol. Six microlitres of RNA was added to 0.5 μl 10x hybridization buffer, and 0.5 μl Cy3-labelled extension primer (2 μM). The DNA primer was hybridized at 65 °C for 5 min, before addition of 5 μl KMM-Cy5-mini and extension at 37 °C for 6 min. Samples were analysed as described for fluoro-tREX.

Mini-tREX. This protocol was adapted from Cervettini et al.²². Total tRNA was isolated as described in protocol A and deacylated. One to two micrograms of RNA was diluted into a total volume of 6 μl and added to 0.5 μl 10x hybridization buffer, and 0.5 μl Cy5-labelled extension primer (2 μM). The DNA primer was hybridized at 65 °C for 5 min, before addition of 5 μl KMM-mini and extension at 37 °C for 6 min. Samples were analysed by running a 15% acrylamide 1x TBE gel (running 200 V, 40–80 min).

Bio-tREX. RNA concentrations were adjusted to match the lowest concentration in the samples being compared. Ten microlitres of RNA was added to 2.5 μl 10x hybridization buffer, 11.5 μl water and 1 μl extension primer (2 μM). The DNA probe was hybridized at 65 °C for 5 min, before addition of 25 μl KMM-bio and extension at 37 °C for 6 min. 10 μl of Dynabeads MyOne Streptavidin C1 magnetic beads (Invitrogen) per reaction were washed 3 times with 200 μl washing buffer, resuspended in 50 μl binding buffer; the beads were added to the extension reaction and binding was performed for at least 30 min at 4 °C, with head-over-tail rotation. The supernatant was removed, and the beads were washed 4 times with 200 μl washing buffer. The washed beads were resuspended in 10 μl FLB and heated to 98 °C for 3 min to release the tRNAs. Beads were removed and the supernatant was directly loaded onto a Novex TBE 6 M urea, 10% or 15% PAGE gel (Invitrogen) and run for 36 min in 0.5x TBE at 270 V. Gels were stained with SYBR Gold (Invitrogen) and imaged on an Amersham Typhoon Biomolecular Imager (GE) using the Cy2 emission filter.

Northern blotting. tRNAs were isolated following the general protocol A or B, omitting the oxidation by NaIO_4 . Two to three micrograms of RNA was loaded onto an acidic urea PAGE gel (9% acrylamide (19:1), 100 mM sodium acetate pH 5, 8 M urea) and the gel was run for 12–16 hours, using 100 mM sodium acetate as running buffer, at 6 W constant power. The gel was stained with SYBR Gold (Invitrogen) to identify the tRNAs and an appropriate section of the gel was cut and blotted using iBlot DNA Transfer Stack (Invitrogen) with the iBlot Dry Blotting System. The tRNAs were cross-linked to the membrane (Stratalinker UV Crosslinker 2400), which was subsequently blocked in Ambion ULTRAhyb-Oligo buffer (Invitrogen) for 30 min. The biotinylated DNA probe was added to a final concentration of 0.2 $\mu\text{g ml}^{-1}$ and hybridized overnight at 37 °C and 160 rpm. The membrane was washed 3 times with 20 ml 0.5x TBE buffer and transferred into 15 ml Odyssey blocking buffer for 20 min

before addition of IRDye 800CW Streptavidin (LI-COR) to a final concentration of 0.2 $\mu\text{g ml}^{-1}$. Finally, the membrane was washed 3 times with 20 ml 0.5x TBE and imaged on an Amersham Typhoon Biomolecular Imager (GE) using the IR long range emission filter.

mRNA extraction and oxidation (A). The volumes given are suited for 2 to 3 ml of cell culture and were adjusted proportionally when required. Chemically competent BL21 cells were transformed with a pColE1 plasmid encoding the stmRNA construct, which was under the control of a T7 promoter and T7 terminator, rescued in SOC, shaken at 220 rpm, 37 °C for 1 h, diluted into 2xYT-am and grown overnight. The overnight cultures were diluted in a ratio of 1:20 into 2xYT-am in absence or presence of the ncM and grown at 37 °C, 220 rpm to an OD_{600} of 0.5–0.8. PylRS production was induced by addition of Isopropyl β -D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and cells were grown for 20 min, 220 rpm at 37 °C.

Cells were centrifuged at 4,200 rcf at 4 °C for 12 min, resuspended in 800 μl resuspension buffer, transferred to a 96-well plate and centrifuged at 4,200 rcf, 4 °C for 12 min. Subsequently, the procedure outlined in the user manual of the Agencourt RNAdvance Cell v2 RNA isolation kit (Beckman) was followed. In brief, 200 μl LBE containing 10 μl proteinase K were used to lyse cells at room temperature for 1 h. 244 μl BBC beads were mixed with 266 μl isopropanol, added to the lysate, and the RNA was bound to the beads for 10 min at room temperature. Beads were washed 3 times with 200 μl 80% ethanol, after the final wash the beads were carefully dried, and the RNA was eluted in 80 μl water.

70 μl of the RNA solution was added to 40 μl resuspension buffer and 7.5 μl of 0.1 M NaIO_4 . The oxidation was run on ice for 1 h and quenched with 10.5 μl 100 mM DTT. To the oxidized RNA, a mix of 1.5 μl of 1.6 M Na_2CO_3 and 16.5 μl DNase I buffer (Ambion DNase I, Thermofisher) was added and the samples were resuspended. Subsequently, 18 μl of DNase I was added, and the RNA was incubated at 37 °C for 30 min. The digestion was cooled on ice and 300 μl Agencourt RNAClean XP beads (Beckman) were added and the RNA bound for 10 min at room temperature. The beads were washed 3 times with 80% ethanol. After the final wash the beads were carefully dried, and the RNA was eluted in 25 μl water.

mRNA extraction and oxidation (B). A similar protocol to the one outlined in procedure A was followed, but the RNA was isolated using acid phenol/chloroform extraction. In brief, BL21 *E. coli* harbouring stmRNAs were grown in 5 ml of 2xYT-am in the presence or absence of the ncM until an OD_{600} 0.5–0.9. At this point stmRNA expression was induced by addition of 1 mM IPTG to a final concentration of 1 mM. After 20 min, cells were collected by centrifugation at 4,200 rcf for 12 min at 4 °C. Cell pellets were resuspended in 800 μl resuspension buffer, transferred into 1.5 ml Eppendorf tubes, and centrifuged for 3 min at 4,200 rcf at room temperature. The supernatant was removed, and the pellets were resuspended in 500 μl SLB. 500 μl of acid-phenol was rapidly added and the tubes were vortexed for 1 min and centrifuged for 6 min, 21,000 rcf at room temperature. 450 μl of the aqueous layer was recovered and 50 μl of 2.5 M KCl was added. Acid-phenol chloroform extraction was repeated, retrieving 400 μl of the aqueous layer. 400 μl of chloroform was added and the samples were vortexed for 1 min followed by centrifugation at 21,000 rcf 6 min at room temperature. 300 μl of the aqueous layer was recovered, and chloroform extraction was repeated with 300 μl chloroform. 200 μl of the aqueous phase were transferred to new 1.5 ml Eppendorf tubes and 10 μl of 0.1 M NaIO_4 was added. The oxidation reaction was run for 1 h on ice. 440 μl of ethanol was added and RNA was precipitated at –20 °C for at least 20 min. The RNA was pelleted by centrifugation at 21,000 rcf at 4 °C for 30 min. The supernatant was removed, and the pellets air-dried for 10 min. The RNA was resuspended in 50 μl water. 30 μl of each RNA sample was digested in 120 μl 1x Ambion DNase I buffer including 12 μl Ambion DNase I

enzyme. Samples were purified with 50 µg Monarch RNA clean-up Kit (NEB) and eluted into 20 µl water. We note, that when the RNA is isolated by acid-phenol chloroform extraction, an additional deacylation step in deacylation buffer is required, to measure acylation of stmRNAs with non-α amino acid monomers.

Fluoro-mREX. RNA concentration for all samples were adjusted to match the lowest concentration in the samples being compared. Six to twelve micrograms of RNA was added to a mixture of 1 µl DNA probe (2 µM), 2.5 µl 10x hybridization buffer and water (added to a final volume of 25 µl). The primer was annealed at 65 °C for 5 min. 25 µl KMM-Cy5 was added. The primer was extended for 6 min, 37 °C. Samples were purified using the 10 µg NEB Monarch RNA clean-up Kit (NEB) and eluted in 12 µl water. 12 µl of orange loading dye was added to each sample. Gel electrophoresis was conducted using 1% agarose gels (cast using NorthernMax MOPS running buffer) in NorthernMax MOPS running buffer at 135 V, 42 min. Gels were stained with SYBR Gold (Invitrogen) and imaged on an Amersham Typhoon Biomolecular Imager (GE) using the Cy2 and Cy5 emission filter.

Bio-mREX. The RNA concentration of all samples, after RNA extraction and oxidation (Methods), was adjusted to match the lowest concentration in the samples being compared.

To perform the pulldown, 6–12 µg of RNA was added to a mixture of 1 µl DNA probe (2 µM), 2.5 µl 10x hybridization buffer and water (added to a final volume of 25 µl). The primer was annealed at 65 °C for 5 min. 25 µl KMM-bio were added. The primer was extended for 6 min, 37 °C. Ten microlitres of Dynabeads MyOne *CI* streptavidin beads (Invitrogen) were washed 2 times with washing buffer, and resuspended in 50 µl binding buffer. Resuspended beads were added to the extension reaction, and the biotinylated stmRNAs were bound to the beads for 1 h, 4 °C, with head-over-tail rotation. The beads were washed on a magnetic stand with 3 × 200 µl washing buffer, two times 200 µl AWB, one time 200 µl washing buffer, one time 200 µl water and were finally resuspended in 13 µl of RHM. After the AWB wash and after the final wash with washing buffer, the beads were transferred into new plastic tubes. The primer was annealed at 65 °C for 5 min. 7 µl of RMM was added and the RNA reverse transcribed at 50 °C for 10 min. One microlitre of RNase H was added and the mixture heated to 37 °C for 15 min and 98 °C for 3 min to release the cDNA from the beads. Finally, the cDNA was separated from the beads and either used for quantification by qPCR, NGS, or as a template for further cloning.

To determine the number of molecules in the input for the stmRNA pulldown, extracted and oxidized RNA samples, in 13 µl of RHM, were reverse transcribed using the same procedure as for the pulled-down stmRNA. The percentage of the input stmRNA molecules recovered in the pulldown was determined from the number of molecules before and after the pulldown, as determined by qPCR ($(\text{number of molecules after pulldown} / \text{number of molecules in input}) \times 100$).

qPCR of cDNA from bio-mREX. qPCR reactions were run in triplicate for each bio-mREX sample and were composed of 2 µl of cDNA, 10 µl PowerUp SYBR Green Master Mix (Applied Biosystems), 0.4 µl of each primer and 7.2 µl water. A standard was generated by PCR of the *MmPylRS* gene and quantified using a Qubit 2 Fluorometer (Life Technologies) and the Qubit 1x dsDNA HS Assay Kit (Invitrogen). A five-step fivefold serial dilution was used to generate a qPCR standard curve. This allowed calculation of qPCR efficiency and the number of molecules in each sample. qPCR was run on a ViiA 7 Real-Time PCR System (Applied Biosystems) using the standard supplier protocol for SYBR Green (Invitrogen).

Preparation of cDNA from bio-tREX for NGS. Half of the cDNA from the 20 µl reverse transcription reaction from bio-tREX was added into a PCR mix containing 25 µl Q5 High-Fidelity 2x Master Mix, 12 µl water and

2 µl of a 10 µM predefined mix of indexing primers (see Supplementary Data 2). A standard PCR program with 29 amplification cycles and an annealing temperature of 60 °C was used. Extension times were adapted to the amplicon length according to the manufacturer's guidelines. DNA was bound to 100 µl of Agencourt AMPure XP (Beckman) for 10 min and the beads were washed 3 times with 200 µl 80% ethanol. Beads were dried and DNA was eluted in 25 µl water. DNA concentrations were measured using Qubit 2 Fluorometer (Life Technologies) and the Qubit 1x dsDNA HS Assay Kit (Invitrogen) and 80 ng of each amplicon were combined into the NGS library. The combined library was diluted in HT1 Hybridization Buffer (Illumina) to a concentration of 2 nM. PhiX (Illumina) was added to increase the diversity of the library at a 20% molar ratio. 12 µl of the library was added to 18 µl HT1 Hybridization Buffer (Illumina) and 20 µl of the diluted mixture was used for NGS analysis.

Cloning of cDNA from bio-tREX for further evolution. Half of the cDNA from the 20-µl reverse transcription reaction, from bio-tREX, was added into a PCR mix containing 25 µl Q5 High-Fidelity 2x Master Mix, 12 µl water and 2 µl of a 10 µM predefined mix of golden gate assembly primers (see Supplementary Data 2). A touchdown PCR program was used. The initial annealing temperature of 65 °C was decreased over 10 cycles by 0.5 °C per cycle. Subsequently, 20 regular cycles using an annealing temperature of 58 °C were performed. Extension times were adapted to the amplicon length according to the manufacturer's guidelines. DNA was bound to 100 µl of Agencourt AMPure XP (Beckman) for 10 min and the beads were washed 3 times with 200 µl 80% ethanol. Beads were dried and DNA was eluted in 25 µl water. The amplicon was then cloned into a new pColE1 backbone, which was previously amplified by Golden Gate primers (Supplementary Data 2), by two-piece Golden Gate assembly according to New England Biolabs' guidelines.

NGS data analysis. NGS was performed on a MiSeq system (in the case of the test evolution with library 1 and substrate 2) or a Next-Seq2000 system (in all other cases). The resulting cDNA from tRNA display was amplified using oligos NGS_A(1-8) and NGS_B(1-8) (Supplementary Data 2) containing different combination of Nextera sequencing barcodes via PCR. Samples were purified, quantified, and combined in equimolar amounts. Paired end reads were first paired using PEAR³⁸, and aligned to a reference sequence of *MmPylRS* using Bowtie2³⁹. The relevant library positions were extracted and translated to amino acids, and resulting variants were counted using R script. Subsequent operations were performed using the frequency of each variant in each library which was computed as the count value divided by the total number of counts of that library. Using R script, enrichment and selectivity scores were calculated for all variants as follows. First, variants that were only present in all positive replicates were considered (tables were merged using AND operator). Assuming that highly enriched sequences could potentially not be covered in the negative and the input samples but may still be of interest, the negative and the naïve replicates were merged to the positive table using an OR operator. A placeholder value of 0.95 counts was adopted in cases where a replicate did not cover a specific variant. The resulting dataset was used to calculate mean enrichments in the presence and in the absence of the nAA or nCM, computed as the quotient of the mean frequencies in one condition and the input condition. The resulting positive and negative enrichments were used to calculate the selectivity value for each variant (equivalent to the quotient of positive and negative frequencies). For further analysis, variants were filtered using an empirically determined threshold value for the normalized standard deviation of the positive frequency (dispersion error in the plus substrate condition).

tRNA pulldown and nCM identification by LC-MS. tRNAs were isolated from 8 ml of cells following the general protocol B omitting the

Article

oxidation by NaIO_4 . The RNA pellet was resuspended in 90 μl buffer D and RNA concentrations adjusted to match the lowest concentration in the samples being compared. A volume of 0.5 μl of biotinylated DNA probe (100 μM) was added to the RNA and the DNA probe was hybridized at 65 °C for 5 min. 40 μl of Streptavidin Dynabeads MyOne C1 (Invitrogen), were washed twice with buffer D-T (buffer D with 0.05% (v/v) Tween), and added into 10 μl buffer D. The beads were added to hybridization reaction and the probe was bound to the beads for minimally 30 min at 4 °C with head-over-tail rotation. The samples were washed 3 times with 200 μl of acW1-T, twice with 200 μl of acW1, 3 times with 200 μl of acW2 and once with 200 μl water, all on a magnetic stand. 24 μl of deacylation buffer was added and the beads, which were incubated at 42 °C for one hour. 12 μl of the deacylation mix was added to 3 μl 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC, 3 mg ml^{-1} in acetonitrile) and the reaction incubated at 55 °C for 15 min. Samples were analysed on an Agilent Technologies 6130 Quadrupole LC-MS using single ion monitoring.

GFP(150X)-His₆ and GFP(3X)-His₆ fluorescence assay. Chemically competent DH10 β cells harbouring a p15A plasmid encoding GFP(150TAG)His₆ or GFP(3TAG)His₆ (two versions of the p15A plasmid with either a tetracycline or apramycin resistance cassette, which led to similar results, were used interchangeably) were transformed with a pMB1 plasmid encoding *Mm*PyIRS, or a mutant thereof, and *M. mazei* tRNA^{PyI}_{CUA}, rescued in SOC and grown overnight in 2xYT_s-t or 2xYT_s-ap. 20 μl of the overnight culture was diluted into 480 μl 2xYT_s-t or 2xYT_s-ap containing 0.2 L-arabinose in presence and absence of 2 to 4 mM of the respective ncM in a 96-well-plate format. Cells were grown for 16–20 h at 37 °C at 700 rpm. The plates were centrifuged for 12 min, 4,200 rcf at 4 °C and the cells resuspended in 150 μl PBS. One-hundred microlitres of the resuspended cells were transferred into a Costar 96-well flat bottom plate and the OD₆₀₀, and GFP fluorescence was measured using a PHERAstar FS plate reader.

GFP(150X)_{His6} isolation for mass spectrometry analysis. Three replicates of the protein produced as described above were combined in a 1.5 ml Eppendorf tube, centrifuged at 4,200 rcf for 3 min, frozen at -20 °C, thawed and resuspended in 150 μl BugBuster (Millipore). Cells were lysed for 1 h with head-over-tail rotation. Lysed cells were centrifuged for 20 min, 20 000 rcf, at 4 °C and the lysate was added to 20 μl of Ni-NTA beads. GFP(150X)-His₆ was bound to the beads for 20 min at room temperature with head-over-tail rotation. The beads were washed 6 times with 60 μl 30 mM imidazole in PBS, and the protein was eluted with 5 times 30 μl 300 mM imidazole in PBS.

For lower-activity mutants for ncM 13, 5–15 ml of cell culture was used for protein production. The volumes of BugBuster were adjusted proportionally, all other volumes were kept the same.

Mass spectrometry. ESI-MS and MS/MS were performed as previously described^{29,33}.

Protein expression, purification and crystallization. Chemically competent DH10 β cells harbouring a p15A plasmid encoding GFP(150TAG) His₆ and a pMB1 plasmid encoding *Mm*PyIRS(13_1) and *Mmt*RNA^{PyI} were transformed, rescued in SOC and grown overnight in 2xYT_s-ap. 10 ml of the overnight culture was diluted into 1 l 2xYT_s-ap containing 0.2% L-arabinose in presence of 5 mM 13.

Bacterial pellet of 1 l expression culture of GFP(150(S) β^3 mBrF)-His₆ was lysed by sonication, centrifuged at 142,000 rcf for 30 min and supernatant bound to Ni-NTA beads (Qiagen). Beads were washed three times before protein was eluted and further purified by gel filtration using a Superdex 75 HiLoad 26/60 pg column (GE Healthcare) in 25 mM Tris pH 7.4, 200 mM NaCl and 0.06% NaN₃. The purified protein was concentrated using Vivaspin 20, 10,000 MWCO (Sartorius) to

6 mg ml^{-1} . Sample was Trypsin digested with Sequencing Grade Modified Trypsin (Promega) in a 50:1 ratio. Sample was incubated for 1 h at 37 °C, centrifuged at 21,000 rcf for 10 min before plating in crystal trays. Crystallization trials with multiple commercial crystallization kits were performed in 96-well sitting-drop vapour diffusion plates (Molecular Dimensions) at 18 °C and set up with a Mosquito HTS robot (TTP Labtech). Drop ratios of 0.2 μl protein solution plus 0.2 μl reservoir solution were used for screening. The only useful dataset was collected from a crystal collected from the Fusion screen (Molecular Dimensions)⁶⁰ with following composition: 37.5% PEG 3350/PEG 1 K/MPD (1:1:1), 0.1 M Bicine/Trizma pH 8.5, 0.8% (w/v) Morpheus III Alkaloids and 0.12 M Morpheus Alcohols. Crystals were collected and flash frozen in liquid nitrogen.

Diffraction data collection, processing and structure solution.

Diffraction data were collected at the ESRF on beamline ID23-2 at an energy of 14.2 keV. Data were processed with XDS⁶¹ via the pipeline autoProc (Global Phasing)⁶². The structure was solved by molecular replacement with MolRep⁶³ using the homologue model PDB 2B3P. Iterative building was performed with Coot⁶⁴, refinement with REFMAC5⁶⁵, and validation with Molprobt⁶⁶. Figures of the structure were prepared with PyMOL (PyMOL Molecular Graphics System, Schrödinger).

Selection for ncAAs. The selection was performed as described in Supplementary Fig. 11. RNA was isolated and oxidized as described in general procedure A. Bio-mREX was performed as specified in the general procedure. After the first round of selection, the new libraries were assembled from the amplified cDNA as described above. After the second round of selection the NGS samples were prepared from the isolated cDNA as described above, the NGS was run using a P2 600 cycles cartridge, and the data were analysed as specified above.

Selection for ncMs. The selections were performed as described in Supplementary Figs. 22 and 35. RNA was isolated and oxidized as described in general procedure A. Bio-mREX was performed as specified in the general procedure. The NGS samples were prepared from the isolated cDNA as described above, the NGS run using a P1 600 cycles cartridge, and the data were analysed as specified above.

Selection for substrate 13 using a random mutagenesis library. The concentrations of the pMB1 plasmids encoding PyIRS hits 13_1 and 13_2 were measured by Qubit 2 Fluorometer (Life Technologies) and the Qubit 1x dsDNA HS Assay Kit (Invitrogen) and the plasmids combined in equimolar amounts. The combined plasmids were used for an error prone PCR of the active site of PyIRS using golden gate primers (Supplementary Data 2) and the GeneMorph II kit (Agilent) at conditions leading to the maximal number of random mutations. The amplicons were cloned into a new pColE1 backbone by two-piece Golden Gate assembly according to NEB (New England Biolabs) guidelines. The selection was performed as outlined in Supplementary Fig. 33. RNA was isolated and oxidized as described in general procedure A. Bio-mREX was performed as specified in the general procedure. The NGS samples were prepared from the isolated cDNA as described above, the NGS was run using a P2 600 cycles cartridge, and the data were analysed as specified above.

Statistics. Graphpad Prism version 9 was used to generate all bar graphs in this study. For sequence alignments and further processing of the NGS data custom R scripts were used.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Primary data for main and supplementary figures are available in Supplementary Data 1. A list of plasmids and oligonucleotides used in this study is available in Supplementary Data 2. Sequencing data generated in this study were deposited at the Sequencing Read Archive under the BioProject ID PRJNA1014447. The structure of GFP(150(S) β 3mBrF)-His6 is available in the Protein Data Bank under accession code 8OVY. All other datasets generated and/or analysed in this study are available from the corresponding author on reasonable request. All materials (Supplementary Data 2) from this study are available from the corresponding author on reasonable request.

Code availability

All code used for analysis of tRNA display data is available at GitHub (<https://github.com/JWChin-Lab/tRNA-display>).

54. Tang, S. et al. Mechanism-based traps enable protease and hydrolase substrate discovery. *Nature* **602**, 701–707 (2022).
55. Neumann, H. et al. A method for genetically installing site-specific acetylation in recombinant histones defines the effects of H3 K56 acetylation. *Mol. Cell* **36**, 153–163 (2009).
56. Stemmer, W. & Morris, S. Enzymatic inverse PCR: a restriction site independent, single-fragment method for high-efficiency, site-directed mutagenesis. *Biotechniques* **13**, 214–220 (1992).
57. Brubaker, L. & McCorquodale, D. The preparation of amino acid-transfer ribonucleic acid from *Escherichia coli* by direct phenol extraction of intact cells. *Biochim. Biophys. Acta* **76**, 48–53 (1963).
58. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
60. Gorrec, F. & Bellini, D. The FUSION protein crystallization screen. *J. Appl. Crystallogr.* **55**, 310–319 (2022).
61. Kabsch, W. Automatic indexing of rotation diffraction patterns. *J. Appl. Crystallogr.* **21**, 67–72 (1988).
62. Vonrhein, C. et al. Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. D* **67**, 293–302 (2011).
63. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66**, 22–25 (2010).
64. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
65. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
66. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).

Acknowledgements This work was supported by the Medical Research Council (MRC), UK (MC_U105181009 and MC_UP_A024_1008) and an ERC Advanced Grant SGCR, all to J.W.C. For the purpose of Open Access, the MRC Laboratory of Molecular Biology has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. D.L.D. was supported by the Boehringer Ingelheim Fonds and Magdalene College, Cambridge. A.D. was supported by the Swiss Study Foundation. The authors thank S. Oehm and Z. Trimov for useful discussions; W. Schmied for initial discussions on acylation-based synthetase selection; C. Franco, H. Kramer, T. Morgan and F. Begum at the MRC-LMB mass spectrometry facility and K. Heesom at the proteomics facility of the University of Bristol for performing mass spectrometry; and A. Gautam for assistance and support in using the beamline. We acknowledge the European Synchrotron Radiation Facility (ESRF) for provision of synchrotron radiation facilities.

Author contributions J.W.C. set the direction of research. D.L.D., C.P. and J.W.C. designed the project. D.L.D. and C.P. led the experiments and data analysis presented in this work. All authors, except J.W.C., performed experiments. M.F. and D.B. contributed to crystallization and crystallography. D.L.D., C.P. and J.W.C. wrote the paper with input from all authors.

Competing interests J.W.C. is a founder of Constructive Bio. The Medical Research Council have filed a patent application based on this work. The other authors declare no competing interests.

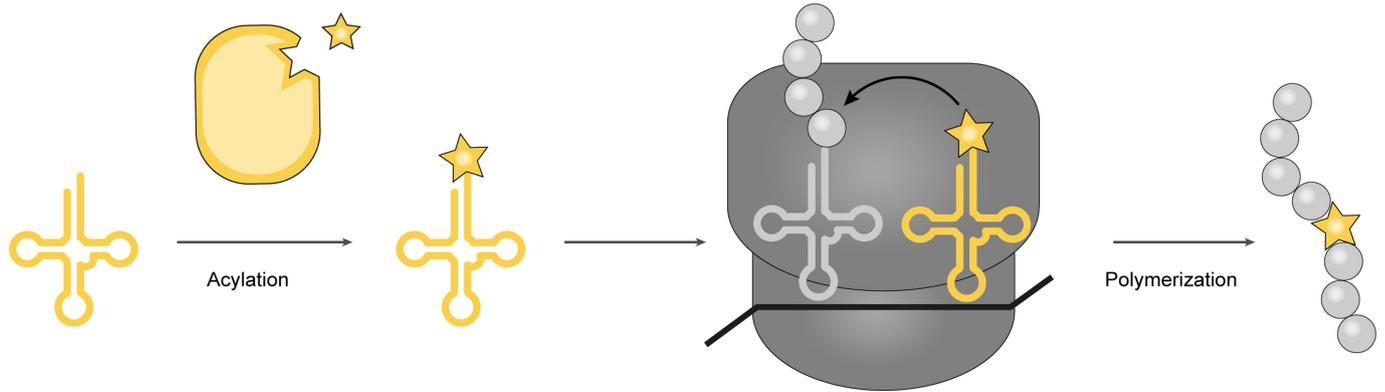
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-023-06897-6>.

Correspondence and requests for materials should be addressed to Jason W. Chin.

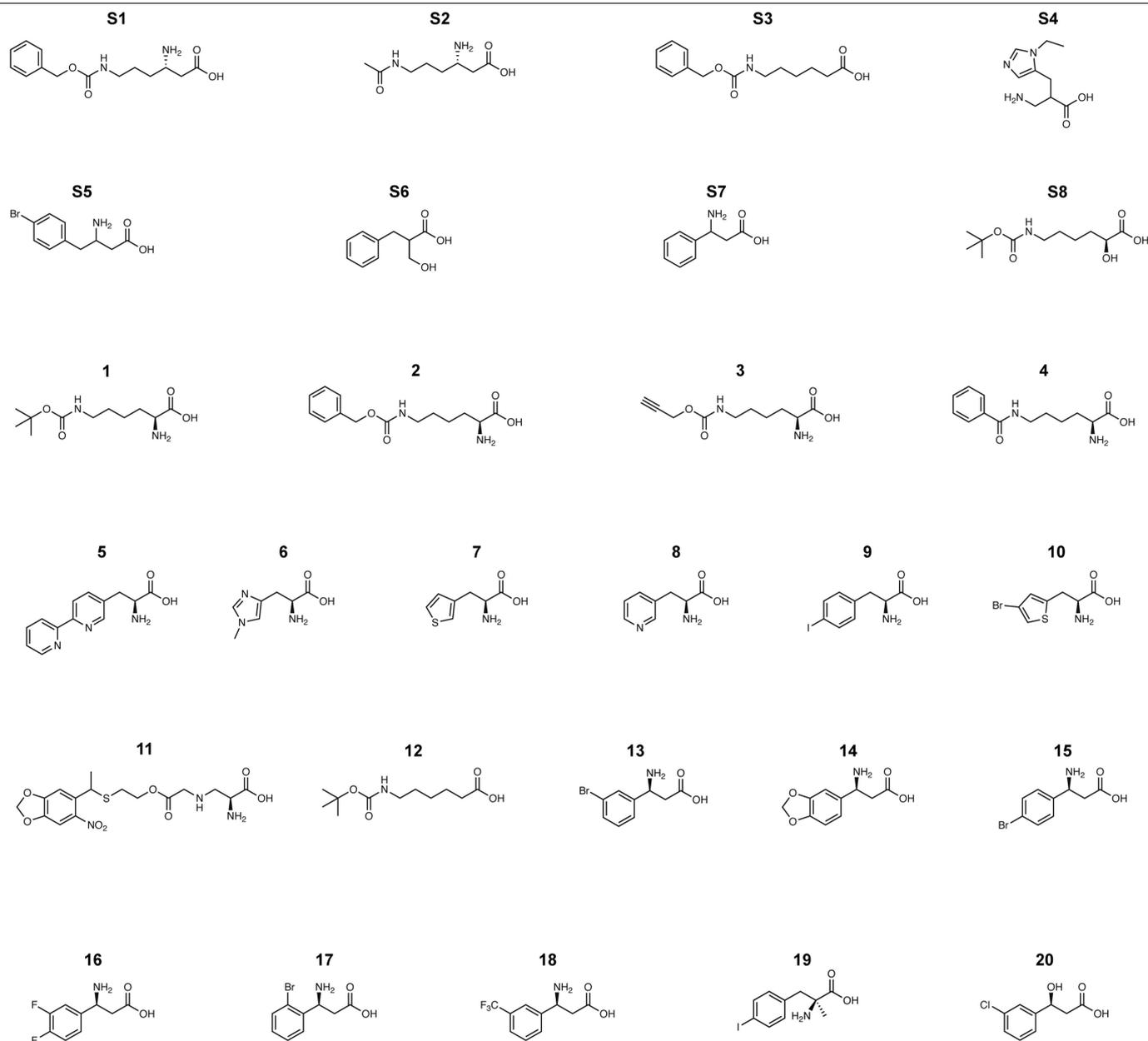
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Encoded cellular incorporation of non-canonical monomers into proteins and into non-canonical polymers requires both tRNA acylation and ribosomal polymerization. The encoded, site specific, incorporation of a non-canonical monomer (ncM, yellow star) via cellular translation requires both the acylation of an orthogonal tRNA with the ncM by an orthogonal synthetase, and ribosomal polymerization of the ncM into a polymer chain; arrow indicates peptide bond formation between A-site monomer and P-site nascent chain. Current methods for engineering aminoacyl-tRNA synthetases that acylate new monomers rely on translational readouts

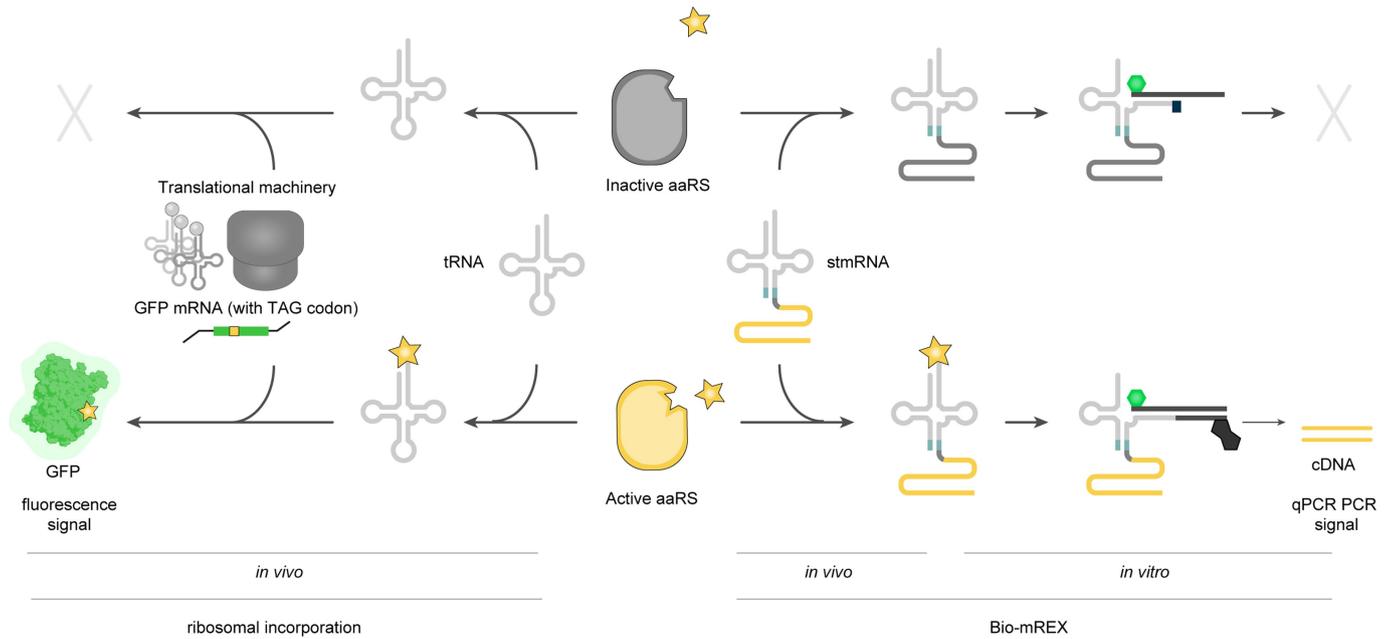
and therefore require the monomers to be ribosomal substrates. For ncMs that are poor ribosomal substrates this co-dependence creates an evolutionary deadlock in cells; an orthogonal synthetase cannot be evolved to acylate an orthogonal tRNA with ncMs that are poor ribosomal substrates, and ribosomes cannot be evolved to polymerize ncMs that cannot be acylated onto orthogonal tRNAs. To break this deadlock, we develop direct selections for orthogonal synthetases to aminoacylate their cognate orthogonal tRNAs with ncMs, independent of whether the ncMs are ribosomal substrates.



Extended Data Fig. 2 | Chemical structures of all compounds used in this study.

(**S1**), (*S*)-3-amino-6-(((benzyloxy)carbonyl)amino)hexanoic acid ((*S*)- β^3 CbzK) (**S1**), (*S*)-6-acetamido-3-aminohexanoic acid ((*S*)- β^3 AcK) (**S2**), 6-(((benzyloxy)carbonyl)amino)hexanoic acid (CbzAhx) (**S3**), 3-amino-2-((1-ethyl-1*H*-imidazol-5-yl)methyl)propanoic acid (β^2 NeH) (**S4**), 3-amino-4-(4-bromophenyl)butanoic acid (β^2 pBrhF) (**S5**), 2-benzyl-3-hydroxypropanoic acid (β^2 OH-F) (**S6**), 3-amino-3-phenylpropanoic acid (β^3 F) (**S7**), (*S*)-6-((*tert*-butoxycarbonyl)amino)-2-hydroxyhexanoic acid (OH-BocK) (**S8**). *N*⁶-((*tert*-butoxycarbonyl)-*L*-lysine (BocK) (**1**), *N*⁶-((benzyloxy)carbonyl)-*L*-lysine (CbzK) (**2**), *N*⁶-((prop-2-yn-1-yloxy)carbonyl)-*L*-lysine (AlkyneK) (**3**), *N*⁶-benzoyl-*L*-lysine (BenzK) (**4**), 3-((2,2'-bipyridin)-5-yl)-2-aminopropanoic acid (BiPyA) (**5**), *N*¹-methyl-*L*-histidine (*N*¹mH) (**6**), (*S*)-2-amino-3-(thiophen-3-yl)propanoic acid (3-ThiA)

(**7**), (*S*)-2-amino-3-(pyridin-3-yl)propanoic acid (PyA) (**8**), (*S*)-2-amino-3-(4-iodophenyl)propanoic acid (pIF) (**9**), (*S*)-2-amino-3-(4-bromothiophen-2-yl)propanoic acid (BrThiA) (**10**), (2*S*)-2-amino-3-(((2-((1-(6-nitrobenzo[*d*][1,3]dioxol-5-yl)ethyl)thio)ethoxy)carbonyl)amino)propanoic acid (pcDAP) (**11**), *N*⁶-((*tert*-butoxycarbonyl)amino)hexanoic acid (BocAhx) (**12**), (*S*)-3-amino-3-(3-bromophenyl)propanoic acid ((*S*)- β^3 mBrF) (**13**), (*S*)-3-amino-3-(benzo[*d*][1,3]dioxol-5-yl)propanoic acid ((*S*)- β^3 MDF) (**14**), (*S*)-3-amino-3-(4-bromophenyl)propanoic acid ((*S*)- β^3 pBrF) (**15**), (*S*)-3-amino-3-(3,4-difluorophenyl)propanoic acid ((*S*)- β^3 pmFF) (**16**), (*S*)-3-amino-3-(2-bromophenyl)propanoic acid ((*S*)- β^3 oBrF) (**17**), (*S*)-3-amino-3-(3-(trifluoromethyl)phenyl)propanoic acid (**18**) ((*S*)- β^3 mCF₃F), (*S*)-2-amino-3-(4-iodophenyl)-2-methylpropanoic acid (**19**) ((*S*)- α -Me-pIF), (*S*)-3-(3-chlorophenyl)-3-hydroxypropanoic acid (**20**) (OH-*S*)- β^3 mClF).



Extended Data Fig. 3 | Relationship between cDNA retrieved from bio-mREX and fluorescence obtained from in vivo genetic code expansion. Relationship between the acylation signal measured, by bio-mREX, for stmRNAs and the GFP fluorescence signal measured for intact, translation-competent tRNAs. For stmRNAs, active aminoacyl-tRNA synthetases (aaRS) lead to the acylation of their encoding stmRNAs, which by bio-mREX get extended, separated and ultimately reverse transcribed. This results in the cDNA of the active synthetase, which can be quantified by qPCR. In the case of an inactive

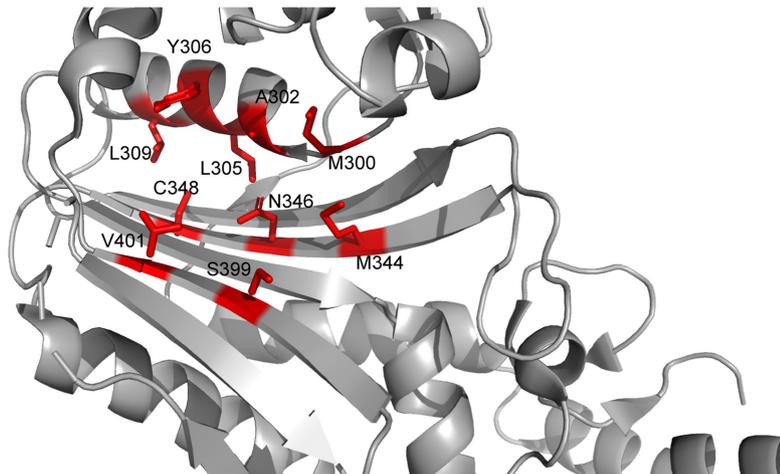
aaRS the stmRNAs is not acylated and no cDNA is produced in bio-mREX experiments. Therefore, the activity of a synthetase in bio-mREX correlates with the number of cDNA molecules measured by qPCR. In canonical translation an active aaRS enzyme leads to an acylated, intact, cognate tRNA_{CUA} which is used in protein translation. Inactive aaRS enzymes lead to non-acylated tRNAs, which are not used in protein translation. The production of GFP protein from *GFP(1SOTAG)His₆*, as measured by GFP fluorescence, reports on the acylation of tRNA_{CUA}, as well as the other steps in the production of protein.

a

<i>M. mazei</i> PyIRS residue	300	302	305	306	309	344	346	348	384	399	401	Theoretical Diversity
1				●	●		●					3.3×10^4
2	●	●		●	●		●					3.4×10^7
13				●	●	①	●	●			●	4.1×10^7
14	②	●	●	●	●	①	③	●			④	1.7×10^8
3D			●		●	①	⑤	●	⑥	⑦	●	1.1×10^8
4D			●	●		①	⑤	●	⑥	⑦	●	1.1×10^8

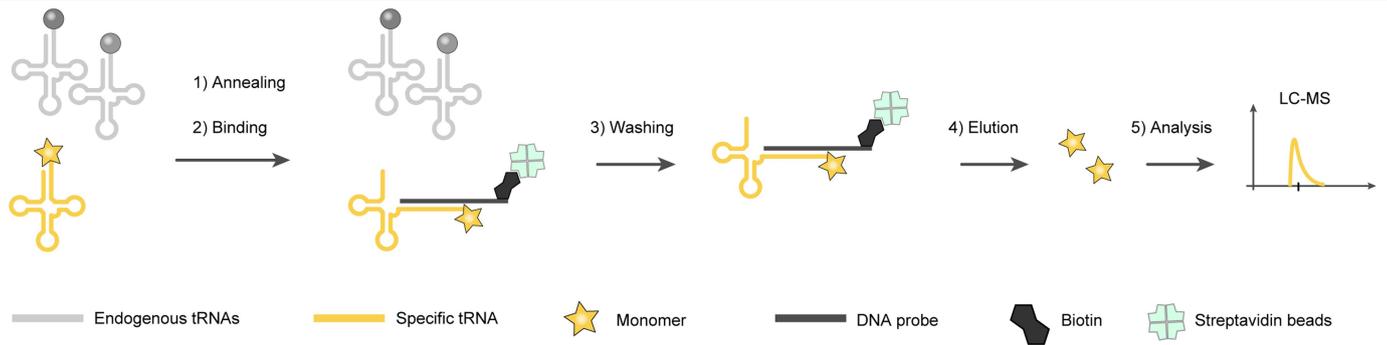
Codon	aa Diversity	Encoding residues
● NNK	(20)	A G L I V F M Y W N Q D E C S T H R K P
● DBK (+AAA)	(13)	A G L I V F M W C S T R K
● NDT	(12)	G L I V F Y N D C S H R
● NRT	(8)	G Y N D C S H R
① Custom Mix 1	(2)	M Q
② Custom Mix 2	(8)	A M N D E C S T
③ Custom Mix 3	(8)	A G V N E C S T
④ Custom Mix 4	(8)	A L V F C S T K
⑤ Custom Mix 5	(9)	A G V N Q E C S T
⑥ Custom Mix 6	(3)	L F Y
⑦ Custom Mix 7	(2)	S T

b



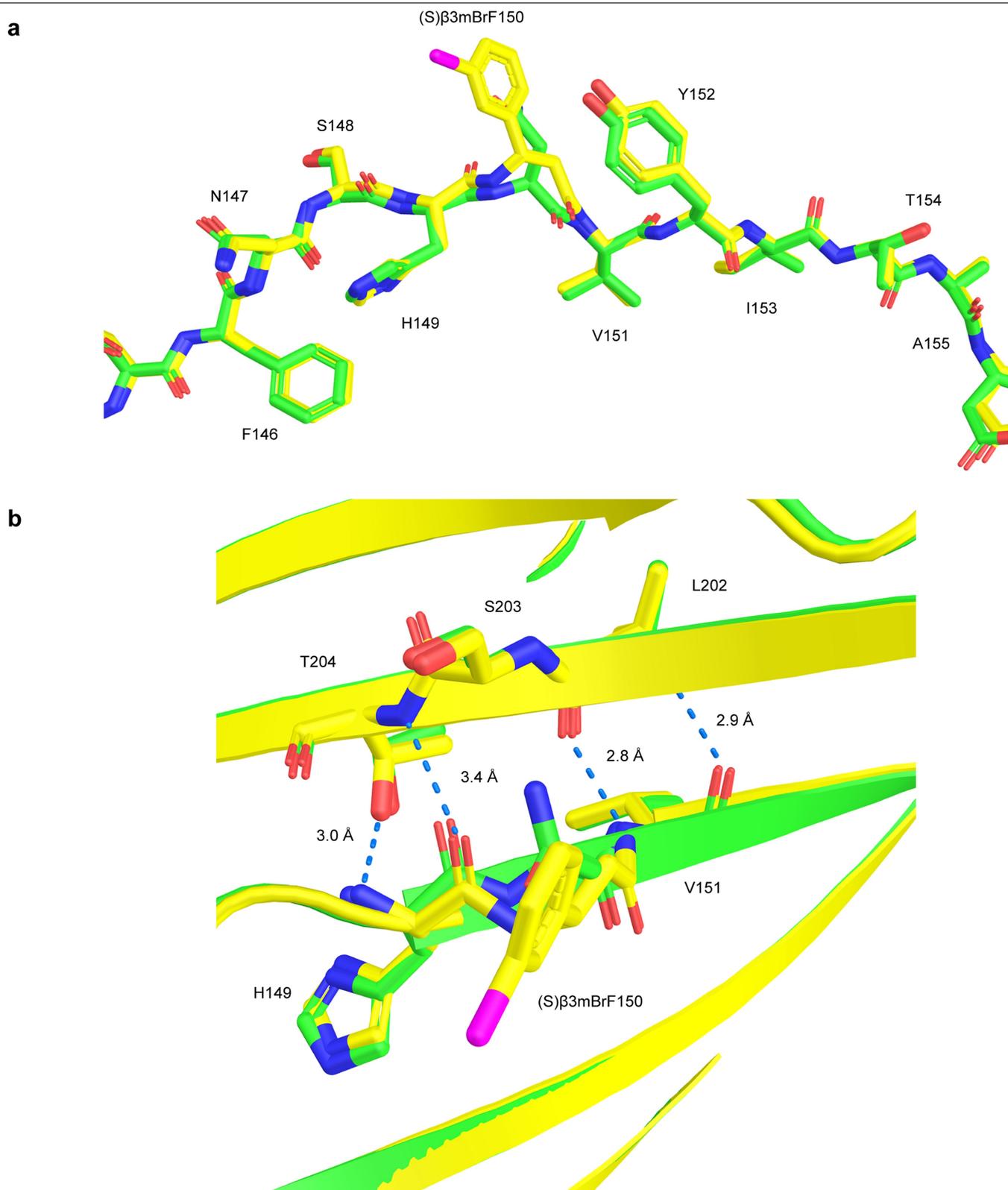
Extended Data Fig. 4 | Library design for this study. PyIRS libraries used in this work. **a**, Overview of the seven libraries designed and created. These libraries target a total of 11 amino acid residues in the PyIRS active site and employed several types of degenerate codons. NNK codons are depicted as dark red, DBK codons (+lysine codon) as blue, NDT codons as dark green, NRT codons as yellow. For certain sites, custom residue mixes encompassing the

most commonly observed mutations were used (1-7 mixes, depicted as grey spheres). All libraries were created with at least 10^9 independent transformants. N = A, T, G, C; K = G, T; D = G, A, T; B = G, T, C; R = G, A. The custom mixes are described in the methods. **b**, The eleven amino acid residues targeted for mutagenesis in the PyIRS active site are shown in red. Image was rendered using Pymol, based on the PDB structure 2ZIN.



Extended Data Fig. 5 | Assay for determination of acylating monomer identity. Schematic of tRNA pulldown followed by LC-MS analysis to determine the identity of the monomer on the target tRNA. tRNAs are extracted from cells expressing the tRNA of interest and the cognate orthogonal aaRS, grown in the

presence of the ncM. A biotinylated probe is annealed, and the targeted tRNA is pulled down. After washing, the ncM is eluted by alkaline deacylation, derivatised with 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC), and detected using LC-MS.



Extended Data Fig. 6 | Detail of the structure of GFP containing a β -amino acid. Details of structure GFP(150(S) β^3 mBrF)-His₆ (PDB code 8OVY). **a**, Detail of protein chain at position 150. The acquired structure for GFP(150(S) β^3 mBrF)-His₆ (in yellow) is superimposed with the structure of wt GFP used for refinement (PDB: 2B3P), showing the kink in the backbone induced by the β^3 -amino acid.

b, Detail of hydrogen bond network in the beta-barrel at position 150. The GFP(150(S) β^3 mBrF)-His₆ structure (yellow) is superimposed to the structure of WT GFP (PDB: 2B3P). The kink induced by incorporation of (S) β^3 mBrF (**13**) affects the local hydrogen bond at that position, however the remaining contacts of the corresponding beta-strand remain intact.

Article

Extended Data Table 1 | Data collection and validation statistics (Molecular Replacement) for the crystal structure of GFP150(S) β^3 mBrF_{His6}

GFP150(S) β^3 mBrF _{His6}	
Data collection	
Wavelength (Å)	0.87313
Space group	C2 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	130.5, 37.7, 91.8
α , β , γ (°)	90, 106.4, 90
Resolution (Å)	45.35 - 1.54 (1.595 - 1.54)
Total reflections	197662 (8327)
Unique reflections	59288 (5803)
R _{merge}	0.11 (1.35)
I/ σ (I)	5.1 (1.4)
Completeness (%)	93.1 (95.1)
Redundancy	3.3 (2.7)
Wilson B-factor	17.6
Refinement	
<i>R</i> _{work} / <i>R</i> _{free}	0.19 / 0.22
Number of non-hydrogen atoms	
Protein	3558
Ligand/ion	68
Water	159
Average <i>B</i> -factors	
Protein	20.3
Ligand/ion	16.2
Water	22.5
R.m.s. deviations	
Bond lengths (Å)	0.016
Bond angles (°)	1.46
Validation	
Ramachandran favored (%)	98.17
Ramachandran allowed (%)	1.60
Ramachandran outliers (%)	0.23
Rotamer outliers (%)	0.00
Clashscore	5.47

The experiment was executed on a single crystal. Statistics for the highest-resolution shell are shown in parentheses.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

- | | |
|-----------------|---|
| Data collection | Illumina sequencer (MiSeq / NextSeq 2000) was used to acquire all sequencing data in this study; raw data has been deposited at the Sequencing Read Archive under the BioProject ID PRJNA1014447. |
| Data analysis | <p>Graphpad Prism version 9 was used to generate all bar graphs in this study. For sequence alignments and further processing of the NGS data custom scripts were used - all software used is listed in the methods section. Custom scripts available on GitHub, link provided in Code Availability Statement.</p> <p>All code used for analysis of tRNA display data is available at GitHub (https://github.com/JWChin-Lab/tRNA-display) [The link will be activated shortly before publication].</p> |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Primary data for main and supplementary figures is available in Supplementary Data 1. A list of plasmids and oligonucleotides used in this study is available in Supplementary Data 2. Sequencing Data generated in this study were deposited at the Sequencing Read Archive under the BioProject ID PRJNA1014447. The structure of GFP(150S) β 3mBrFHis6 is available in the Protein Data Bank under accession code 8OVY. All other datasets generated and/or analysed in this study are available from the corresponding author on reasonable request. All materials (Supplementary Data 2) from this study are available from the corresponding author on reasonable request.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. These are biochemical experiments, not animal experiments, and there are therefore no individuals to sample. Sample size is not a relevant parameter, but the number of replicates is relevant. The number of replicates is indicated in the relevant figure legend.
Data exclusions	No data was excluded.
Replication	All fluorescence measurements presented in the main text were carried out in independent triplicates or quadruplets and little variation was observed. All fluorescence measurements were reproducible. All attempts at replication were successful.
Randomization	These are biochemical experiments, where different components need to be added to different reactions. A single investigator needs to perform defined, distinct and skilled operations on different samples to make the experiment meaningful and therefore randomization does not make sense.
Blinding	These are biochemical experiments, where different components need to be added to different reactions. A single investigator needs to perform defined, distinct and skilled operations on different samples to make the experiment meaningful and therefore blinding does not make sense.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions	<i>Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).</i>
Location	<i>State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).</i>
Access & import/export	<i>Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).</i>
Disturbance	<i>Describe any disturbance caused by the study and how it was minimized.</i>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	<i>Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Validation	<i>Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.</i>

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	<i>State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology and Archaeology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No | Yes

- | | | |
|--------------------------|--------------------------|----------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> | Public health |
| <input type="checkbox"/> | <input type="checkbox"/> | National security |
| <input type="checkbox"/> | <input type="checkbox"/> | Crops and/or livestock |
| <input type="checkbox"/> | <input type="checkbox"/> | Ecosystems |
| <input type="checkbox"/> | <input type="checkbox"/> | Any other significant area |

Experiments of concern

Does the work involve any of these experiments of concern:

- | No | Yes | |
|--------------------------|--------------------------|---|
| <input type="checkbox"/> | <input type="checkbox"/> | Demonstrate how to render a vaccine ineffective |
| <input type="checkbox"/> | <input type="checkbox"/> | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| <input type="checkbox"/> | <input type="checkbox"/> | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| <input type="checkbox"/> | <input type="checkbox"/> | Increase transmissibility of a pathogen |
| <input type="checkbox"/> | <input type="checkbox"/> | Alter the host range of a pathogen |
| <input type="checkbox"/> | <input type="checkbox"/> | Enable evasion of diagnostic/detection modalities |
| <input type="checkbox"/> | <input type="checkbox"/> | Enable the weaponization of a biological agent or toxin |
| <input type="checkbox"/> | <input type="checkbox"/> | Any other potentially harmful combination of experiments and agents |

Plants

- | | |
|-----------------------|--|
| Seed stocks | <i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i> |
| Novel plant genotypes | <i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i> |
| Authentication | <i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i> |

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

- | | |
|--|--|
| Data access links
<i>May remain private before publication.</i> | <i>For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.</i> |
| Files in database submission | <i>Provide a list of all files available in the database submission.</i> |
| Genome browser session
(e.g. UCSC) | <i>Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.</i> |

Methodology

- | | |
|-------------------------|--|
| Replicates | <i>Describe the experimental replicates, specifying number, type and replicate agreement.</i> |
| Sequencing depth | <i>Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.</i> |
| Antibodies | <i>Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.</i> |
| Peak calling parameters | <i>Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.</i> |
| Data quality | <i>Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.</i> |
| Software | <i>Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.</i> |

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI

Used

Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.***Statistical modeling & inference**

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*Specify type of analysis: Whole brain ROI-based Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.(See [Eklund et al. 2016](#))

Correction

*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).***Models & analysis**

n/a | Involved in the study

 Functional and/or effective connectivity Graph analysis Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.