Discovery of widespread activating mutations in a compact RNA-guided endonuclease

Brittney W. Thornton[1,2†], Rachel F. Weissman[1,2†], Ryan V. Tran[3,4], Brenda T. Duong[3], Jorge E. Rodriguez[1,2], Cynthia I. Terrace[1,2], Evan D. Groover[5,2], Jung-Un Park[1,2,6], Julia Tartaglia[1,2], Jennifer A. Doudna[1,2,3,6,7,8,9], David F. Savage[1,2,6*]

[†]These authors contributed equally to this work.

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, 94720, USA
[2]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, 94720, USA
[3]Department of Chemistry, University of California, Berkeley, Berkeley, CA, 94720, US
[4]Scribe Therapeutics, Alameda, CA, 94501, USA
[5]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA
[6]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, 94720, USA
[7]California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, CA, US
[8]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA
[9]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

*Corresponding author, Email: savage@berkeley.edu

Abstract

TnpB is an evolutionarily diverse family of RNA-guided endonucleases associated with prokaryotic transposons. Due to their small size and putative evolutionary relationship to Cas12s, TnpB holds significant potential for genome editing and mechanistic exploration. However, most TnpBs lack robust gene-editing activity, and unbiased profiling of mutational effects on editing activity has not been experimentally explored. Here, we mapped comprehensive sequence-function landscapes of a TnpB ribonucleoprotein and discovered many activating mutations in both the protein and RNA. Single-position changes in the RNA outperform existing variants, highlighting the utility of systematic RNA scaffold mutagenesis. Leveraging the mutational landscape of the TnpB protein, we identified enhanced protein variants from a combinatorial library of activating mutations, which increased editing levels in yeast and human cells relative to wild-type TnpB. In total, this study highlights unknown elements critical for

regulation of endonuclease activity in both the TnpB protein and the RNA, and reveals a surprising amount of latent activity accessible through mutation.

Introduction

TnpBs are a family of RNA-guided endonucleases encoded within IS200/IS605 and IS607 transposons and are thought to be the evolutionary ancestors of Type V CRISPR-Cas enzymes[1–5]. TnpB binds to the right end element RNA (reRNA), which has a noncoding 5' scaffold and a 3' variable region, to guide ribonucleoprotein (RNP) cleavage at a complementary DNA sequence proximal to a transposon-associated motif (TAM)[2,3]. As a putative evolutionary predecessor to the CRISPR-Cas12 enzymes, TnpB retains core domains shared across this CRISPR protein family[4–7]. Understanding the relationship between TnpB's protein sequence and biochemical function can provide both fundamental knowledge and a basis for engineering improved or altered RNA-guided RuvC endonuclease activities.

While highly active variants of CRISPR-Cas enzymes have been discovered through protein engineering and rational design[8], these approaches often explore a limited sequence space. The conformational changes TnpB undergoes during the dynamic coordination of nucleic acid binding, catalytic center activation, and DNA cleavage[6,7] make it challenging to predict the effects of mutations.

Deep mutational scanning (DMS) approaches typically assay every individual amino acid mutation, using high-throughput assays to assess protein function[9–11]. While DMS effectively provides comprehensive maps of protein function, it is often practically limited by protein size. With a compact amino acid length and an RNA scaffold, TnpB is well-suited for this approach. We conducted DMS over the entire ISDra2 TnpB RNP, utilizing a positive selection assay for DNA cleavage to identify a broad spectrum of enhancing, neutral, and deleterious mutations within the TnpB protein and its reRNA. These data elucidate the dynamic regions involved in DNA binding and cutting, including a mutational hot-spot within the reRNA secondary

structure where mutations increase DNA cleavage activity. We found that 20% of single amino acid mutations, many of which are not frequently observed in nature, increase activity relative to the wild-type (WT) TnpB protein. This finding suggests that native ISDra2 TnpB activity may be subject to negative selection, possibly due to its role as a transposon-associated homing endonuclease[2,12]. Furthermore, we identified combinations of activating mutations that increase TnpB-mediated genome-editing activity in human cells.

Results and Discussion

Selection for TnpB-mediated DNA cleavage in yeast

In their native genomic context, the TnpB reRNA and protein coding sequences overlap with each other and with insertion sequence elements essential for transposition, imposing unknown evolutionary constraints[2,3]. To interrogate the effects of mutations in both reRNA and TnpB protein without the native sequence constraints of transposition, we encoded the reRNA and codon-optimized ISDra2 TnpB protein under the control of separate regulatory elements (Fig. 1a). To assess on-target cleavage activity we adapted an *in vivo* selection previously used to enhance CRISPR-Cas9 activity, which utilizes yeast (*S. cerevisiae*) strains with a genomic *ade2*$^-$ reporter cassette[13] (Fig. 1b). On-target cleavage of the reporter cassette initiates *ADE2* repair, enabling cells to grow on media lacking adenine. Thus, reporter strain growth in the presence and absence of adenine can be used as a readout for target site cleavage (Fig. 1c). We first demonstrated that this assay can quantitatively measure endonuclease activity beyond that of CRISPR-Cas9, including that of WT ISDra2 TnpB and CRISPR-Cas12 endonucleases[13] (Extended Data Fig. 1a).

We constructed separate pooled plasmid libraries of reRNA and protein variants (Fig. 1a). The reRNA and protein DMS libraries were barcoded such that each variant was associated with ~30 unique barcodes to provide statistical replicates. The libraries were transformed into yeast reporter strains with either their WT reRNA or protein

counterparts, and barcode abundance in selective and nonselective media was quantified at multiple timepoints across two biological replicates. Relative variant enrichment was calculated as the log ratio of variant abundance in selective and non-selective conditions, and all enrichments were normalized to wild-type controls (Fig. 1c).

Profiling the mutational landscape of the TnpB reRNA

The reRNA accounts for nearly half of the molecular weight of the ISDra2 TnpB ribonucleoprotein (RNP) complex[6,7]. Previously, a truncation within the reRNA's stem 2 region, termed Trim2, has been shown to maintain or increase ISDra2 TnpB's genome editing activity[7,14] (Extended Data Fig. 1b). Given the complex secondary and tertiary structures of the TnpB RNA scaffold, we hypothesized that comprehensive mutagenesis of the reRNA scaffold might provide insight into evolutionary constraints on the reRNA sequence and RNP function.

The 116 nt reRNA scaffold is necessary and sufficient for TnpB-mediated cleavage[7] and was chosen as the starting sequence for deep mutational scanning (Extended Data Fig. 1c). To investigate the mutational tolerance of the reRNA, our DMS library covered all single-nucleotide substitutions and single- and double-nucleotide deletions. Additionally, we included variants with disordered regions of stem 1 and stem 2 replaced with thermodynamically stable tetraloops[15,16], the reported Trim2 variant, and variants with reported inactivating truncations within the triplex and pseudoknot as negative controls[14], for a total of 576 assayed mutants (Fig. 2a-b).

Upon mapping the reRNA mutational landscape, we found that inactivating truncations and deletions within the pseudoknot were depleted as expected under selective conditions (Fig. 2b). Stable tetraloop replacements in stem 2 were more highly enriched above WT than those in stem 1, and Trim2 was one of the most highly enriched variants. Unexpectedly, we observed substitutions and deletions that were more enriched than both WT and the Trim2 variant (Fig. 2b). These activating mutations

were concentrated around unpaired nucleotides rA^{-40}-rU^{-43} within stem 2 (Fig. 2a,c). We hereafter refer to this region (rA^{-37}-rU^{-44}; rG^{-75}-rG^{-79}) as the "hinge" region of the reRNA, which appears to create a sharp bend in stem 2 preceding the disordered distal end[6] (Fig. 2d). We also assessed hinge deletion variants in HEK293T cells using an enhanced green fluorescent protein knockout (EGFP KO) assay. When targeted to an EGFP transgene, reRNA variants resulted in an increase in EGFP KO compared to the WT reRNA, as assessed with flow cytometry[17] (Fig. 2e-f).

Stem 2 has been proposed by Sasnauskas *et al.* to serve as a regulatory switch that controls the transition of the TnpB RNP into a cleavage-competent conformation upon DNA binding and heteroduplex formation[6]. This activation is driven by a conformational change of stem 2, where the formation of the RNA-DNA heteroduplex displaces stem 2's distal end, leading to the release and activation of the RuvC domain. We hypothesize that the activating hinge mutations enhance TnpB activity by increasing the flexibility of the distal end of stem 2, making it more prone to displacement. This may facilitate the release of the RuvC domain, making it easier to be activated. While the details of mechanistic insights underlying this increased activity remain to be elucidated, this hypothesis is consistent with the previous report that stem 2 truncation leads to dysregulated collateral ssDNA cleavage, independent of target DNA binding[6].

Deep Mutational Scanning of TnpB Protein

We next mapped the fitness landscape of the ISDra2 TnpB protein with a library spanning all possible single amino acid substitutions and stop codons, alongside catalytically inactive (dead) and WT protein controls (Fig. 3). We collected data on 93% (7,611 of 8,140) of all possible substitutions across two biological replicates, which were highly reproducible with an R^2 of 0.81 (Fig. 4a). We found that 3.7% of these substitutions were enriched by at least two-fold compared to WT TnpB (Figs. 3 and 4a).

Enrichment density varied across domains, with many mutations enriched in the RuvC and WED domains, most ZnF mutations depleted, and mutations in the unstructured C-terminal tail largely neutral (Fig. 3, Extended Data Fig. 2a). Stop codons that would cause truncations were depleted except for those at the C-terminus following residue 376, consistent with previous *in vitro* data showing that the C-terminal tail is dispensable for target cleavage[7] (Extended Data Fig. 2b). Additionally, we observed depletion of alanine substitutions at residues important for recognition of the TAM, which is essential for cleavage at the adjacent target sequence[2,3,7] (Fig. 4a).

Positively charged amino acids were enriched over negatively charged residues within the vicinity of nucleic acids, particularly within the central channel, where the TAM-proximal end of the heteroduplex is accommodated[6] (Extended Data Fig. 3). This is consistent with previous findings in CRISPR Cas12 enzymes, where mutations to positively charged amino acids near the guide RNA-DNA heteroduplex, close to the protospacer adjacent motif (PAM), have been shown to increase activity and affect specificity[18,19].

Similarly, specific WED-domain residues, N4 and L172, which stabilize the first TAM-proximal base pair of the heteroduplex,[6] were enriched for aromatic and small functional groups, respectively (Fig. 4b). We reason that the initial guide-target duplex formation may be enhanced by small hydrophobic and nucleophilic amino acids at position 172, and by π-stacking interactions between aromatic residues at N4 and the first heteroduplex nucleobases. The introduction of aromatic amino acids near the first base-pair of the RNA-DNA heteroduplex has also been associated with increased activity in engineered Cas12i2, PlmCas12x, and AsCas12f variants[20,21]. These findings provide evidence for a shared activation mechanism within an interaction conserved across TnpB and CRISPR-Cas endonucleases.

Within the positively charged central channel, E302 was highly enriched for substitutions to any other amino acid that was not negatively charged (Fig. 4b). Positioned near the heteroduplex backbone, E302 might lead to electrostatic repulsion

with the target strand (TS) phosphate backbone, potentially reducing RNP activity. We also found other mutational hotspots, such as I304, where substitutions were enriched for residues having a range of physicochemical properties.

Unexpectedly, hydrophobic amino acids were enriched at position P282, which lies at the boundary of the lid subdomain that blocks the RuvC active site from accessing the TS[6] (Figs. 3 and 4b). The lid subdomain makes non-sequence-specific contacts with the heteroduplex minor groove, which may aid in sensing heteroduplex formation prior to RuvC activation. We hypothesize that the substitution of wild-type proline residue at this position to small, hydrophobic residues may increase the flexibility of the lid subdomain, accelerating the conformational change required to sense heteroduplex formation prior to TnpB activation.

Overall, 844 mutations from the DMS dataset were enriched over WT TnpB with a p-value <0.05, and were distributed across both nucleic-acid binding interface and the protein surface (Fig. 4c). We tested twenty highly enriched mutants in HEK293T cells with the EGFP KO assay. All reduced EGFP expression, with P282I resulting in a nearly four-fold reduction relative to WT TnpB (Fig. 4d). The discovery of several activating mutations and mutational hotspots suggests that TnpB activity may be natively subject to negative selection.

To explore the generalizability of this protein DMS dataset, we transferred pairs of activating mutations to TnpB orthologs ISYmu1 and ISAba30, which share significant structural similarity with ISDra2 TnpB (0.93 and 0.92 TM-Scores for ISYmu1 TnpB and ISAba30 TnpB, respectively), despite low sequence similarity (59% and 48% for ISYmu1 and ISAba30)[22,23] (Extended Data Fig. 4a). We designed two ISYmu1 TnpB variants (H4Y/V305R and L167G/V305R) and one ISAba30 TnpB (L4Y/V272I) variant by introducing pairs of analogous activating mutations from ISDra2 TnpB (N4Y/I304R; L172G/I304R; and N4Y/P282I, respectively). We observed increased colony reversion with all three variants compared to their WT orthologs in the yeast cleavage assay (Extended Data Fig. 4b-c). These results demonstrate that these mechanisms of

activation are generalizable beyond ISDra2 TnpB, highlighting the potential for untapped genome-editing activity.

Combinatorial mutations enhance TnpB activity

To explore increases in TnpB editing activity through mutation combinations, we selected 33 highly enriched single amino acid mutations covering 19 positions across TnpB (Fig. 3). Using nicking mutagenesis, we generated a library of ~5×10^3 variants with an average of ~5 of the 33 possible mutations per variant[24,25] (Extended Data Fig. 5a). The combinatorial variant library underwent selection in two separate reporter yeast strains with different target sequences (Fig. 5a, Extended Data Fig. 5b). We observed the greatest increase in enrichment in variants with 4-5 mutations on average, while depleted variants had more variable mutation numbers (Extended Data Fig. 5c). The expression levels of eight highly active TnpB variants were found to be similar to wild-type by Western blot, consistent with a change in enzymatic activity rather than protein abundance (Extended Data Fig. 6).

To assess the genome-editing activity of TnpB combinatorial variants, we selected six of these variants (eTnpB1a–eTnpB1f) for evaluation at four endogenous genomic loci in human cells (Fig. 5b). HEK293T cells were transfected with plasmids encoding individual TnpB protein variants targeting endogenous loci, and genomic DNA was harvested four days post-transfection. When compared to WT TnpB, all six variants demonstrated higher indel (insertion/deletion) formation frequencies across all four target sites, with the exception of eTnpB1c at the *AGBL1* target site. Indel frequencies ranged from 15-56% with eTnpB1f, while eTnpB1d demonstrated the highest overall indel formation frequencies across multiple loci (23-42%), surpassing both wild-type ISDra2 and ISYmu1 TnpB (11%-29% and 6%-30%, respectively) (Extended Data Fig. 7a).

To evaluate off-target activity, we identified six genomic sites with 4–6 mismatches to the target-complementary reRNA sequence with Cas-OFFinder[26]. All

variants exhibited increased off-target indel formation compared to wild-type ISDra2 and ISYmu1 with at least two sites, with up to 6% off-target indel formation frequencies observed with eTnpB1c at *TET1* off-target site 1 (Extended Data Fig. 7b). We generally observed lower indel frequencies at off-target sites with more TAM-proximal mismatches preceding the 12th nucleotide of the heteroduplex. These observations align with published data that TAM-distal mismatches are more well-tolerated by WT TnpB, as only the first 12 TAM-proximal base-pairs of the heteroduplex form contacts with the protein in the ternary complex[6,7,27]. While increases in on-target editing were consistently accompanied by increases in off-target indel frequencies, eTnpB1e was associated with the lowest off-target activity of the variants (<2% at all sites). These findings suggest that some combinations of activating mutations may impair TnpB fidelity, highlighting a potential evolutionarily disadvantageous tradeoff between on and off-target endonuclease activity that will be important to explore in future work.

We investigated whether genome editing activity could be further enhanced by combining highly active TnpB protein and reRNA variants (Extended Data Fig. 7c). We used the EGFP-KO assay to test five TnpB protein mutants paired with one of two reRNA mutants (ΔU^{-42} or ΔC^{-74}-G^{-75}), each with deletions on opposite strands of the hinge region. However, a number of the tested pairings with the reRNA mutants did not lead to substantial additive improvements in editing, relative to pairings with the WT reRNA. Instead, combining the highly active eTnpB1d variant with either hinge deletion variant resulted in reduced genome editing compared to the wild-type reRNA (15-24% vs 44.5% $EGFP^-$ cells when assessed as described above). One possible explanation is that certain protein-reRNA variant combinations destabilize the TnpB RNP beyond a critical free energy threshold, disrupting RNP assembly or dsDNA targeting[28]. Further studies are needed to clarify how protein and reRNA mutations interact to influence RNP stability and function. Systematic combinatorial testing of reRNA and protein variants may reveal more optimal pairings that enhance RNP stability and activity.

Conclusions

In the *Deinococcus radiodurans* genome, ISDra2 TnpB is encoded alongside the HUH superfamily TnpA transposase, where both rely on overlapping sequences essential for transposition and endonuclease activity[2,29]. The reRNA stem 1 sequence overlaps with the imperfect hairpin in the transposon right end (RE), which is required for TnpA recognition and excision[6,29,30]. Notably, stem 1 exhibited a high degree of mutational tolerance, as did the few protein residues within the vicinity of stem 1, supporting the hypothesis that TnpB protein and reRNA have co-evolved with TnpA and the transposon[31] (Extended Data Figs. 3a and 8).

The activity of TnpB is further confounded by the selective balance of transposon maintenance, transposon spread, and host toxicity[12]. By profiling TnpB-mediated on-target cleavage outside the context of transposition, we identified many activating mutations across the RNP, highlighting the rugged nature of the mutational landscape[7,32]. The frequency of activating and non-deleterious mutations in TnpB is an outlier compared to common models of protein evolution and other mutational studies[28,33–35]. This aligns with the hypothesis that TnpB exhibits pervasive evolutionary flexibility, having been exapted for diverse biological processes across multiple clades of life[1,31,36,37]. Additionally, the prevalence of evolutionarily-accessible activating mutations may suggest TnpB endonuclease activity is under negative selection pressure in the transposon context.

This work presents the first comprehensive sequence-function landscapes for both the protein and RNA scaffold of an RNA-guided endonuclease. Comprehensive reRNA mutagenesis uncovered an unexpected mutational hotspot in stem 2, and offers an alternative approach to iterative reRNA and gRNA engineering through truncations and G:U swaps to optimize gene-editing activity[6,21,38]. Mutational scanning of the TnpB protein not only reproduced published findings on ISDra2 TnpB point mutants[7,32], but also captured many novel activating mutations that increased on-target cleavage

activity. We further demonstrate that highly activating mutations can be combined to enhance genome-editing activity in HEK293Ts, and we present eTnpB1a-eTnpB1e as enhanced TnpB variants. Further exploration with alternative delivery methods and broadening of the 5'-TTGAT-3' TAM recognition motif will increase the utility of TnpB for genome-editing applications. Overall, these comprehensive mutagenesis libraries provide molecular insights into nucleic acid binding, activation, and cleavage by TnpB, mapping both mutational constraints and activating mutations across the ribonucleoprotein. Further biochemical and epistatic studies may help elucidate mechanisms of activating mutations in TnpB and related endonucleases. In addition to laying the groundwork for further engineering, we hope our findings will provide insights into TnpB evolution and TnpB's function within insertion sequences.

References

1. Altae-Tran, H. *et al.* Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2308224120 (2023).
2. Karvelis, T. *et al.* Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).
3. Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
4. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).
5. Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
6. Sasnauskas, G. *et al.* TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature* **616**, 384–389 (2023).
7. Nakagawa, R. *et al.* Cryo-EM structure of the transposon-associated TnpB enzyme. *Nature* **616**, 390–397 (2023).
8. Koonin, E. V., Gootenberg, J. S. & Abudayyeh, O. O. Discovery of diverse CRISPR-Cas systems and expansion of the genome engineering toolbox. *Biochemistry* **62**, 3465–3487 (2023).
9. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
10. Prywes, N. *et al.* A map of the rubisco biochemical landscape. *Nature* 1–6 (2025).
11. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160**, 882–892 (2015).

12. Meers, C. *et al.* Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863–871 (2023).
13. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).
14. Li, Z. *et al.* Engineering a transposon-associated TnpB-ωRNA system for efficient gene editing and phenotypic correction of a tyrosinaemia mouse model. *Nat. Commun.* **15**, 831 (2024).
15. Antao, V. P., Lai, S. Y. & Tinoco, I., Jr. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **19**, 5901–5905 (1991).
16. Banáš, P. *et al.* Performance of molecular mechanics force fields for RNA simulations: Stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.* **6**, 3836–3849 (2010).
17. Chen, K. *et al.* Lung and liver editing by lipid nanoparticle delivery of a stable CRISPR-Cas9 RNP. *bioRxiv* (2023) doi:10.1101/2023.11.15.566339.
18. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.* **37**, 276–282 (2019).
19. Zhang, H. *et al.* An engineered xCas12i with high activity, high specificity, and broad PAM range. *Protein Cell* **14**, 538–543 (2023).
20. Chen, Y. *et al.* Synergistic engineering of CRISPR-Cas nucleases enables robust mammalian genome editing. *Innovation (Camb)* **3**, 100264 (2022).
21. Hino, T. *et al.* An AsCas12f-based compact genome-editing tool derived by deep mutational scanning and structural analysis. *Cell* **0**, (2023).
22. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure

template quality. *Proteins* **57**, 702–710 (2004).

23. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
24. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).
25. Mighell, T. L., Toledano, I. & Lehner, B. SUNi mutagenesis: Scalable and uniform nicking for efficient generation of variant libraries. *PLoS One* **18**, e0288158 (2023).
26. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
27. Marquart, K. F. *et al.* Effective genome editing with an enhanced ISDra2 TnpB system and deep learning-predicted ωRNAs. *Nat. Methods* 1–10 (2024).
28. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
29. Hickman, A. B. *et al.* DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. *EMBO J.* **29**, 3840–3852 (2010).
30. Chandler, M. *et al.* Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat. Rev. Microbiol.* **11**, 525–538 (2013).
31. Yoon, P. H. *et al.* Eukaryotic RNA-guided endonucleases evolved from a unique clade of bacterial enzymes. *Nucleic Acids Res.* **51**, 12414–12427 (2023).
32. Cheng, P. *et al.* Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell Res.* (2024) doi:10.1038/s41422-024-00989-2.

33. Bershtein, S. & Tawfik, D. S. Ohno's model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts. *Mol. Biol. Evol.* **25**, 2311–2318 (2008).

34. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).

35. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Advances in Neural Information Processing Systems* **36**, 64331–64379 (2023).

36. Wiegand, T. *et al.* TnpB homologues exapted from transposons are RNA-guided transcription factors. *Nature* (2024) doi:10.1038/s41586-024-07598-4.

37. Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023).

38. Xu, X. *et al.* Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol. Cell* **81**, 4333–4345.e4 (2021).

39. Ellison, E. E. *et al.* Multiplexed heritable gene editing using RNA viruses and mobile single guide RNAs. *Nature Plants* **6**, 620–624 (2020).

40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

42. Stuckey, S. & Storici, F. Gene knockouts, in vivo site-directed mutagenesis and other modifications using the delitto perfetto system in Saccharomyces cerevisiae. *Methods Enzymol.* **533**, 103–131 (2013).

43. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

44. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

45. Gilchrist, C. L. M., Mirdita, M. & Steinegger, M. Multiple protein structure alignment at scale with FoldMason. *bioRxiv* 2024.08.01.606130 (2024) doi:10.1101/2024.08.01.606130.

46. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).

47. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

48. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224–226 (2019).

49. Baum, P., Thorner, J. & Honig, L. Identification of tubulin from the yeast Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 4962–4966 (1978).

50. He, S. *et al.* The IS200/IS605 family and 'peel and paste' single-strand transposition mechanism. in *Mobile DNA III* 609–630 (American Society of Microbiology, 2015).

Methods

Deep mutational libraries construction

The TnpB reRNA DMS library was constructed from an oligonucleotide pool from Twist Bioscience, and covered the 116 nt reRNA scaffold with flanking primer-binding sites for PCR amplification. The reRNA scaffold library contained ~600 variants, including all nucleotide substitutions; single and double nucleotide deletions; a set of double mutations in the pseudoknot; and stable tetraloop replacements in the disordered reRNA regions. The oligonucleotide library was amplified using KAPA HiFi HotStart ReadyMix with an initial denaturation at 95°C for 3 minutes, 16 cycles of 98°C for 20 seconds, 64°C for 15 seconds, and 72°C for 45 seconds, and a final extension at 72°C for 1 minute. The amplified library was cloned into an intermediate storage vector with NEBuilder HiFi DNA Assembly Master Mix. The reRNA library was then assembled with a destination vector containing the variable 3' reRNA sequence by Golden Gate cloning with BsaI-HFv2 and MlyI.

The reRNA plasmid library was digested with KpnI and ApaI and barcoded by assembly (NEBuilder HiFi DNA Assembly Master Mix) with ssDNA oligonucleotides with internal 15xN barcodes. The barcoded plasmid assembly was transformed into TransforMax EC100D pir-116 Electrocompetent *E. coli* and bottlenecked such that a larger culture for plasmid purification was inoculated with ~2.4×10^4 transformed cells (~40 barcodes x ~600 variants), estimated from colony-forming units (CFUs) counted from titer plates. Control plasmids containing WT and catalytically dead ISDra2 TnpB were barcoded similarly.

The ISDra2 TnpB protein sequence was codon optimized for expression in *S. cerevisiae* and mammalian cells and divided into six segments of 204 bp. For each segment, mutations for all single amino acid changes and stop codons were designed and purchased as oligonucleotide pools from Twist Bioscience with flanking primer binding sites, for a total of 8116 variants. Mutations were designed using the most common *S. cerevisiae* codons, except in cases where this would create a restriction site that would interfere with library cloning. In these cases, an alternative common codon set was used to introduce the intended mutation. The first methionine was excluded from mutagenesis. The six sub-libraries were amplified using KAPA HiFi HotStart ReadyMix with an initial denaturation at 95°C for 3 minutes, 18 cycles of 98°C for 20 seconds, 64-67°C for 15 seconds, and 72°C for 45 seconds, and a final extension at 72°C for 1 minute. Amplified sub-libraries were assembled by Golden Gate cloning with BsaI-HFv2 and six corresponding intermediate cloning vectors containing the flanking WT ISDra2 sequence. Each sub-library plasmid pool was digested with BsmBI-v2, and the concentration of the digested full-length ISDra2 protein coding sequence for each sub-library was measured using a Qubit 4

Fluorometer. Each digested sub-library was mixed at an equimolar ratio and inserted into the destination vector.

Single-stranded 30xN barcode sequences were cloned into the NotI-HF- and XhoI-digested plasmid library with NEBuilder HiFi DNA Assembly Master Mix. Assemblies were transformed into TOP10 cells, and a larger culture for plasmid purification was inoculated with ~$2x10^5$ transformed cells (~24 barcodes x ~8100 variants).

Combinatorial library construction

Two combinatorial libraries with an average of ~3 and ~5 mutations per variant were created using nicking mutagenesis, as previously described[24,25]. DNA oligos covering 19 amino acid positions and 33 possible mutations in the TnpB protein were phosphorylated and pooled in an equimolar ratio. For synthesis of the second strand, 5 pmols and 50 pmols of the phosphorylated oligo pool were initially added with 0.38 fmols of the ssDNA template plasmid, and 4.3 pmols and 43 pmols of the phosphorylated oligo pool were spiked in three times following 5 cycles of amplification, to generate the lower and higher mutation-frequency libraries, respectively.

RRYNx25RY and YYRNx25YR barcodes were cloned into the lower and higher mutation-frequency plasmid libraries, respectively, by assembly with ssDNA oligos, as described in the protein DMS library barcoding. The barcoded library assemblies were bottlenecked to ~$5x10^4$ barcoded variants with an average of ~15 barcodes per variant. The lower and higher mutation-frequency libraries were combined in a 1:12 ratio along with barcoded, catalytically inactivated TnpB protein controls prior to transformation into yeast.

Variant-barcode mapping

After library construction, variants were associated with their barcodes using long-read sequencing (PacBio Sequel II for the protein DMS library and Nanopore MinION for the reRNA DMS and combinatorial libraries). All reads were aligned to a reference plasmid and barcode sequences extracted using Minimap2 and SAMtools[40,41]. Sub-alignments were made for all reads with a given barcode, and a consensus sequence was created using SAMtools for all barcodes with at least two reads for PacBio sequencing and at least 10 reads for nanopore sequencing. Barcodes of incorrect length and consensus sequences containing non-programmed mutations were discarded. For the reRNA, protein, and stacked libraries, 606, 7766, and 6592 variants were mapped respectively, with an average of 33, 28, and 15 barcodes per variant respectively.

Reporter yeast strain creation

Yeast *ade2*$^-$reporter strains were created with the *delitto perfetto* approach[42]. An intermediate *ADE2* knockout was derived from S*accharomyces cerevisiae* BY4741 (ATCC 201388, Meyen ex E.C. Hansen) using the CORE cassette GSKU, excluding Gal-I-SceI, as previously described. To create reporter strains, the intermediate strain was co-transformed with linearized DNA containing the target site flanked by duplicate homology regions, and with plasmid carrying SpyCas9 targeting the CORE cassette. SpyCas9 was constitutively expressed, triggering double-stranded breaks (DSBs) in the CORE cassette and repair with the linear DNA template. Target site integration was confirmed by PCR amplification and sequencing, and the strain was cured of the Cas9 plasmid. Using this approach, we generated *ade2*$^-$ reporter strains UniPAM1, UniPAM2 (target 1 strain), and UniPAM5 (target 2 strain).

Yeast pooled library selection assays

For each experimental replicate, 8-10 µgs of total plasmid containing the DMS or combinatorial libraries of the TnpB protein was transformed. For the reRNA DMS library, 1.5 µgs total of plasmid library was transformed. Yeast were transformed with lithium acetate/single-stranded carrier DNA/PEG method[43].

After transformation, cells were resuspended in synthetic complete drop-out media lacking leucine (SCD-leucine) to select for transformation of plasmids, and recovered overnight at 30°C. The following morning, a fraction of the culture was removed for a pre-induction time point. The remaining cells were induced in leucine liquid media with 2% galactose(w/v) at an initial OD_{600} of 1.0. Induced cultures were allowed to grow at 30°C, and culture samples were removed at multiple time points, pelleted and washed in mQ water, and plated on selective (-adenine -leucine) and non-selective (+adenine -leucine) plates. Several cell concentrations were plated on bioassay dishes (Thermo Fisher) at each time point to ensure maximal library coverage. Before plating, all cultures were grown with supplemental (160mg/ml) adenine. Yeast plates were incubated at 30°C for 48 hours, after which colonies were scraped and plasmid DNA extracted using Zymoprep Yeast Plasmid Miniprep II. Barcodes were amplified from plasmid DNA using KAPA HiFi HotStart ReadyMix with 6-12 cycles for PCR1 and 10 cycles for PCR2. PCRs were cleaned up with Ampure XP beads (Beckman Coulter) and submitted for 150 bp paired-end sequencing on Illumina NextSeq sequencer at the IGI NGS sequencing core.

Variant enrichment calculations

Barcode enrichment was assessed by calculating the log ratio of reads containing a given barcode in selective and nonselective samples. Barcodes with fewer than five reads in selective or non-selective conditions were removed from analysis and the log ratio was normalized by the total number of reads in selective and nonselective

sequencing samples. For the protein and stacked-protein libraries, variant enrichment was calculated as the median barcode enrichment for all barcodes associated with a given variant. For the reRNA library, variant enrichment was calculated as the mean of all barcode enrichments as this produced higher replicate correlation. Variant enrichments were normalized to WT such that WT has an enrichment value of zero. A two-sided Mann-Whitney test was performed to calculate the statistical significance and effect size for each variant for each replicate.

Yeast cleavage assays

To compare the activities of TnpB variants and orthologs, as well as of CRISPR-Cas effectors, yeast cells were transformed with 0.5 µg-1.5 µg clonal plasmids and induced as described above. At the pre-induction and post-induction timepoints, approximately 1-3 OD_{600} units were removed from the transformed yeast culture, washed with mQ water, resuspended in 200 µL mQ water, and serially diluted in triplicate. Serial dilutions were plated on selective (-adenine -leucine) and non-selective (+adenine -leucine) solid media 8 hours post-induction, unless otherwise specified. Plates were incubated at 30°C for 48 hours, after which colony counts from serial dilutions were used to estimate the total number of colony-forming units (CFUs). Colony reversion was calculated by dividing the number of CFUs on selective media over the number of CFUs on nonselective media, multiplied by 100 for percentage.

TnpB protein Western blots in yeast

At 24 hours post-induction of a yeast cleavage assay, 2.5 OD_{600} units of yeast cells were harvested for Western blots at 3000 x *g* for 5 minutes. Cells were resuspended in 100 µl mQ water before being lysed by adding 100 µl 0.2M NaOH and incubating at room temperature for 5 minutes. Cell lysis was pelleted by centrifugation at 21,000 x *g* for 2 minutes, washed with 200 µl of 1X PBS, and pelleted again at 21,000 x *g* for 2 minutes. Pellets were resuspended in 30 µl 1X PBS and 30 µl 4x Laemmli buffer (62.5 mM Tris-HCl, pH 6.8, 10% glycerol, 1% LDS, 0.005% Bromophenol Blue) (BioRad). Samples were boiled at 95°C for 3 minutes before 12 µl of supernatant was loaded for SDS-PAGE on a 4-20% Criterion TGX Precast Midi Protein Gel and separated at 125V for 60 minutes. Transfer to a Bio-Rad Trans-Blot Turbo Midi PVDF Transfer Pack was performed with the Trans-Turbo Turbo Transfer System. The membrane was blocked with 5% milk in TBST for 1 hour at room temperature, and then incubated overnight with mouse anti-FLAG (1:10000, Sigma F1804) and rabbit anti-PGK-1 (1:30000) in 2.5% milk in TBST overnight at 4°C. The membrane was washed three times with TBST for 10 minutes each at room temperature, then incubated with goat anti-mouse and anti-rabbit secondary antibodies (1:10000, LiCor 926-32210) for one hour at room temperature. The membrane was washed again as described before. Images were acquired on the LI-COR Odyssey CLx and processed by using ImageStudio.

Comparative sequence and structure alignments

A multiple sequence alignment (MSA) was created using a previously described pipeline[44]. The ISDra2 TnpB protein sequence was used as the Jackhmmer query for 5 search iterations against the UniRef90 database with a domain and sequence bit score threshold of 0.1. Redundant sequences were removed using HHfilter and sequences with less that 50% coverage were removed. Per position amino acid frequencies were determined by calculating the ratio of the amino acid's occurrence to the total number of sequences with a residue present at that position, as previously described[10].

Structural alignments between ISDra2, ISYmu1, and ISAba30 TnpBs were generated with FoldMason[45] (shown) and pairwise alignments[46] with the AlphaFold2-predicted WT amino acid sequences[47]. Similarity was calculated with a Blosum62 matrix with a threshold of 1[23].

Mammalian genome editing

Mammalian cells (HEK393T or HEK293T-GFP) were grown in Dulbecco's Modified Eagle Medium with high glucose, GlutaMAX supplement, and pyruvate (Thermo Fisher) supplemented with 10% fetal bovine serum (Avantor Seradigm) at 37°C and 5% CO_2. Cells were seeded at approximately 10,000 cells per well in 96-well plates 16-24 hours before transfection. The transfection mix was prepared by combining plasmids encoding the protein and reRNA/sgRNA (100 ng carrying TnpB, 145 ng carrying SpyCas9 for 2.6 fmols/transfection) with 9μL Opti-MEM I Reduced Serum Medium (Thermo Fisher) and 0.3μL TransIT-293 per transfection. Transfection mixes were incubated at room temperature for 30 minutes and added dropwise to the cells.

For flow cytometry, transfected plates were passaged two days post-transfection and then harvested for flow cytometry after two to five days. Cells were trypsinized with 30 μL 0.25% trypsin+EDTA was added to cells for 5 minutes at 37°C, and quenched with 120 μL 1X PBS. Cells were transferred to 96-well round-bottom plates and analyzed by flow cytometry on an Attune NxT Flow Cytometer with an autosampler. Data was analyzed using FlowJo software.

For sequencing, cells were harvested 4 days after transfection, and genomic DNA was extracted with QuickExtract (Lucigen) according to manufacturer instructions. Lysate was used directly for polymerase chain reaction (PCR). PCR products were cleaned with Ampure XP beads (Beckman Coulter), analyzed by a 4150 TapeStation (Agilent), and submitted for 150 bp or 300 bp paired-end sequencing on MiSeq or NextSeq sequencer at the IGI NGS sequencing core. The frequencies of the mutations were assessed by CRISPResso2[48].

Off-target analysis

To assess the specificity of TnpB and TnpB variants, CRISPR RGEN Tools (Cas-OFFinder, http://www.rgenome.net/cas-offinder/) was used to predict potential genomic off-target sites containing the "TTGAT" TAM/PAM and 4-6 mismatches in the spacer sequence[26]. Primers were designed using NCBI Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/).

Acknowledgements

We thank Noam Prywes (University of California, Berkeley, CA) for advice on library design and general guidance on deep mutational scanning and data analysis. We thank Luke Oltrogge (University of California, Berkeley, CA) for key assistance with data analysis and sequencing. We thank Kai Chen (University of California, Berkeley, CA) for providing the HEK293T-EGFP cell line used for EGFP knock-out assays. We thank John Desmarais (CSHL), Brady Cress (University of California, Berkeley, CA), Amy Eggers (Seek Labs), and Joshua Cofsky (Harvard University) for establishing protocols and providing components of expression vectors that we used for the yeast cleavage assay. Antiserum against phosphoglycerate kinase (PGK) was generously provided by Dr. Jeremy Thorner (University of California, Berkeley, CA)[49]. We thank Netra Krishnappa (University of California, Berkeley, CA) for assistance in running NGS samples, and Julia Turnšek (University of California, Berkeley, CA) for help with preliminary protein purification. We additionally thank Owen Tuck (University of California, Berkeley, CA) for insights into the protein DMS dataset and experimental design, and Maria Lukarska (University of California, Berkeley, CA) for reviewing the manuscript.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752 (BWT, RFW, and CIT). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. RVT was funded by the Rose Hills Foundation as part of UC Berkeley's Summer Undergraduate Research Fellowship Program. JP is an HHMI Fellow of The Jane Coffin Childs Memorial Fund. DFS is an Investigator of the Howard Hughes Medical Institute and this research was funded by NIH grant 1R01GM127463. U. S. Department of Energy, Physical Biosciences Program, Award Number DE-SC0016240 (DFS). JAD is an investigator of the Howard Hughes Medical Institute, and research in the Doudna lab is supported by the Howard Hughes Medical Institute (HHMI), NIH/NIAID (U19AI171110, U54AI170792, U19AI135990, U01AI142817), NIH/NINDS (U19NS132303), NIH/NHLBI (R21HL173710), NSF (2334028), DOE (DE-AC02-05CH11231, 2553571, B656358), Lawrence Livermore National Laboratory,

Apple Tree Partners (24180), Rett Syndrome Research Trust, UCB-Hampton University Summer Program, Mr. Li Ka Shing, Koret-Berkeley-TAU, Emerson Collective and the Innovative Genomics Institute (IGI).

Author information

Authors and Affiliations

Innovative Genomics Institute, University of California Berkeley, Berkeley, CA, USA
Brittney W. Thornton, Rachel F. Weissman, Jorge E. Rodriguez, Cynthia I. Terrace, Evan D. Groover, Jung-Un Park, Julia Tartaglia, Jennifer A. Doudna, David F. Savage

Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA, USA
Jung-Un Park, Jennifer A. Doudna, David F. Savage

Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA, USA
Brittney W. Thornton, Rachel F. Weissman, Jorge E. Rodriguez, Cynthia I. Terrace, Julia Tartaglia, Jennifer A. Doudna, David F. Savage

Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA
Evan D. Groover, David F. Savage

Department of Chemistry, University of California Berkeley, Berkeley, CA, USA
Brenda T. Duong, Jennifer A. Doudna

Scribe Therapeutics
Ryan V. Tran

Ethics declarations

Competing interests

BWT, RFW, DFS, and JAD have submitted a related patent.
DFS is a co-founder and scientific advisory board member of Scribe Therapeutics.
The Regents of the University of California have patents issued and pending for CRISPR technologies on which J.A.D. is an inventor. J.A.D. is a cofounder of Azalea Therapeutics, Caribou Biosciences, Editas Medicine, Evercrisp, Scribe Therapeutics and Mammoth Biosciences. J.A.D. is a scientific advisory board member at Evercrisp, Caribou Biosciences, Intellia Therapeutics, Scribe Therapeutics, Mammoth Biosciences, The Column Group and Inari. She also is an advisor for Aditum Bio. J.A.D.

is Chief Science Advisor to Sixth Street, a Director at Johnson & Johnson, Altos and Tempus, and has a research project sponsored by Apple Tree Partners. RVT is an employee of Scribe Therapeutics.

Figures

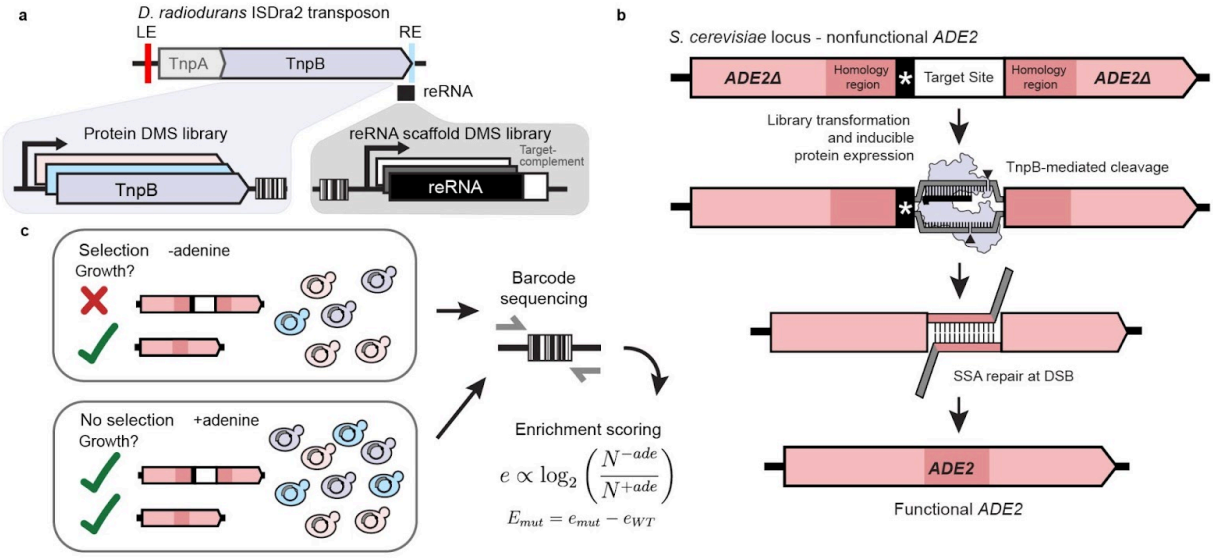


Figure 1. Design of deep mutational scanning libraries and optimized *in vivo* selection for endonuclease activity in yeast.

a, TnpB protein and reRNA from the *D. radiodurans* ISDra2 transposon were placed under the control of separate regulatory elements. Deep mutational scanning libraries were constructed for both molecules and assayed separately. **b,** Schematic of the yeast-based cleavage assay used for individual variant testing and high-throughput library experiments. On-target double-stranded breaks in the reporter cassette enable repair of the *ADE2* locus by single-stranded annealing at duplicate homology regions. *ADE2* repair rescues colony growth in selective, adenine-deficient media. **c,** Representation of the library selection. Plasmid DNA from yeast grown in selective (-adenine) and non-selective (+adenine) conditions was extracted. Barcodes from plasmids were sequenced and log-ratio of barcode abundance in selective over nonselective conditions was used to assess variant enrichment. Enrichment of variants was normalized to the wild-type TnpB RNP enrichment.

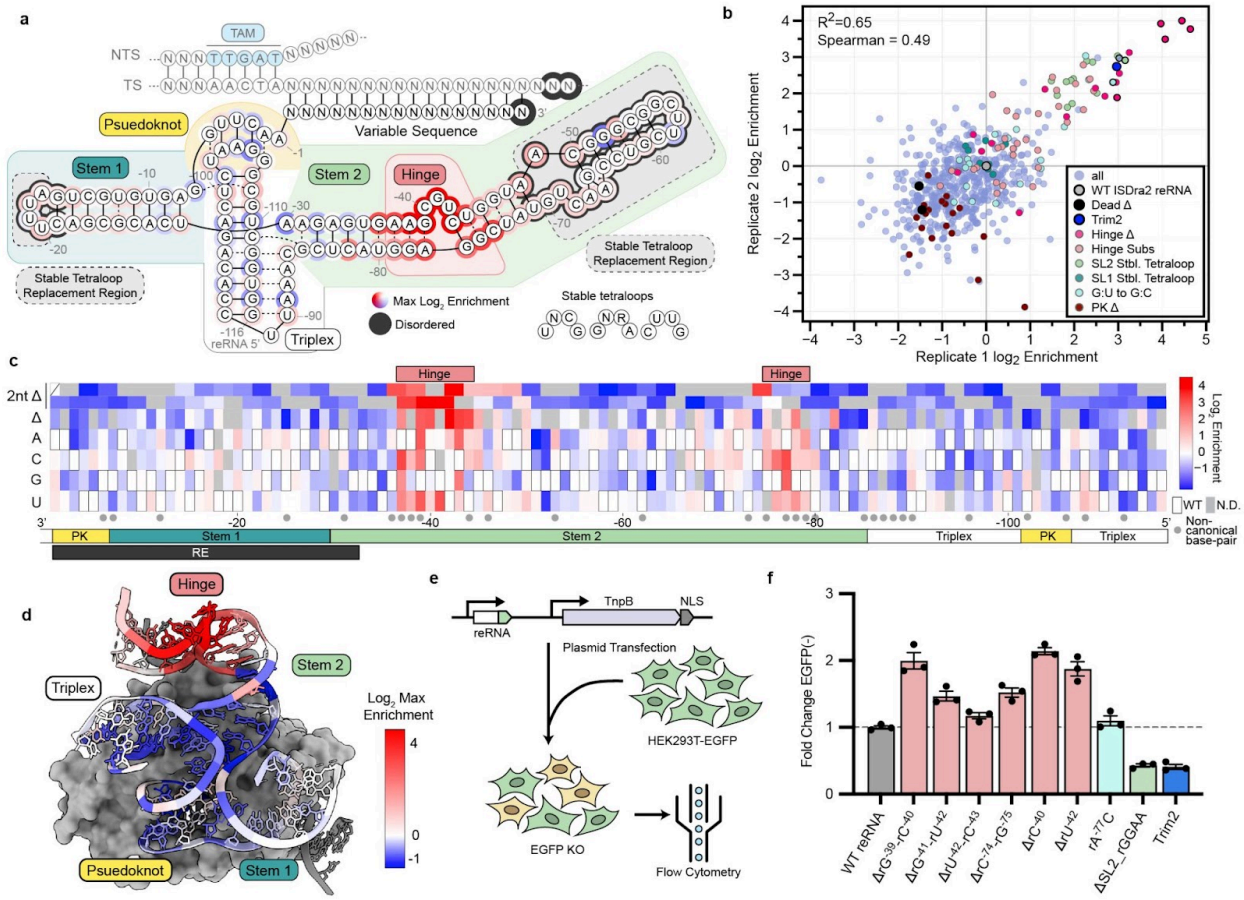


Figure 2. Profiling the TnpB reRNA mutational landscape reveals single nucleotide gain-of-function mutations.

a, Schematic of the ISDra2 reRNA based on ternary RNP cryo-EM structure, adapted from Sasnauskas et al. 2023[6]. Dashed boxes indicate truncated regions replaced with stable tetraloops. Circles around bases are colored by maximum $\log_2$ enrichment scores for a substitution at each position, using the same color scale as in c, Outer gray-circled bases indicate regions not modeled in the ternary structure, designated here as disordered. **b,** Log_2 enrichment of reRNA variants across two experimental replicates. Data is normalized such that WT TnpB is at 0. **c,** Heatmap of enrichment scores for reRNA variants with single-nucleotide substitutions or single- and double-nucleotide deletions. WT positions are indicated with white boxes, and gray boxes denote positions with no available data. **d,** Structure of TnpB ternary complex (PDB ID: 8EXA), with reRNA colored by maximum $\log_2$ enrichment scores at each position **e,** Experimental workflow of EGFP knockout assay in HEK293T EGFP^+ cells. EGFP^- cells were assessed by flow cytometry to compare editing activities of TnpB variants, as described in Methods. **f,** EGFP KO assay in HEK293T EGFP^+ cells for highly enriched reRNA mutants outlined with a bolder black circle in (b). Fold change of

EGFP^{-} percent population for each variant compared to WT TnpB is shown. Data are presented as the mean ± s.e.m. (standard error of mean) from biological replicates (n = 3). Coloring scheme of bars matches the legend in (b).

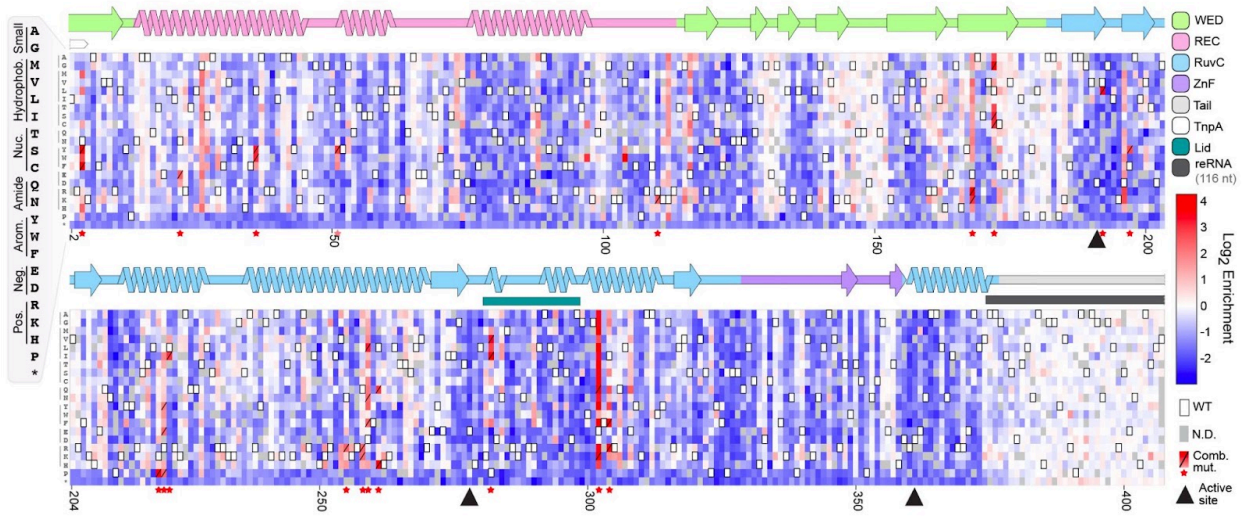


Figure 3. Deep mutational scanning of the TnpB protein identifies mutations that increase activity.

Heatmap showing $\log_2$ enrichment scores of all single amino acid changes. TnpB secondary structures and domains are annotated above the heatmap. White boxes represent wild-type residues. Gray boxes denote positions with no available data. Boxes with a slash and a star at the x-axis represent a mutation that was included in the combinatorial variant library. Black triangles represent active site residues.

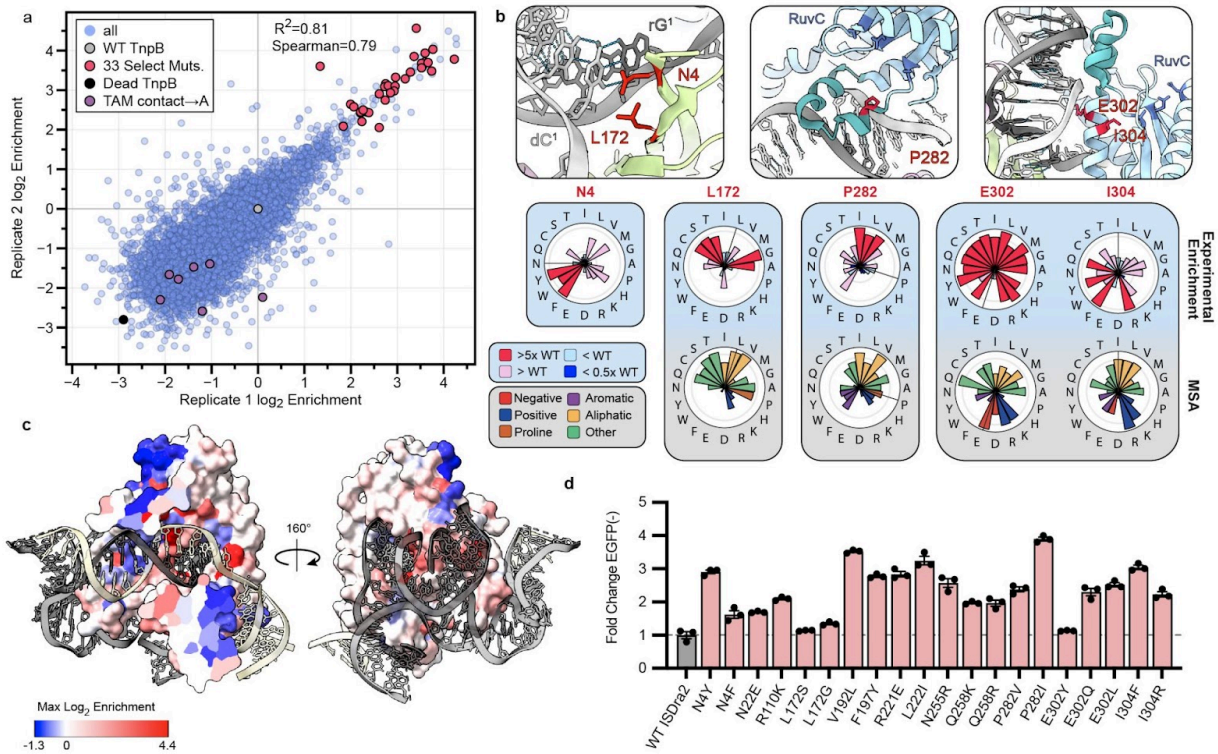


Figure 4. Activating mutations inform mechanistic insights and engineering.

a, Log_2 enrichment of single amino acid variants across two experimental replicates. Data is normalized such that WT TnpB is at 0. Thirty-three mutations were selected from the combinatorial library. Black data points correspond to alanine mutations at the TAM-interacting residues (Y52A, K76A, Q80A, T123A, S56A, F77A, N124A) shown to abrogate or reduce TnpB activity[7]. **b,** Close-up view of residues N4, P282, E302, and I304 in red in the ISDra2 TnpB ternary structure (PDB ID: 8EXA). Site-wise amino acid enrichment and multiple-sequence alignment conservation are displayed on radar plots below. N4 data not shown due to low sequence conservation at the N-terminus. **c,** Max enrichment per residue mapped onto the surface of ISDra2 TnpB ternary structure, shown as a cross-section (left) allowing for visibility of the heteroduplex (PDB ID: 8EXA). reRNA is shown in gray and DNA in light yellow. **d**, Activity of the highly enriched single amino acid mutants was assessed with the EGFP KO assay in HEK293T $EGFP^+$ cells. Fold change of $EGFP^-$ cells (percent of population) for each variant relative to WT TnpB is shown. Data are presented as the mean ± s.e.m. from biological replicates (n = 3).

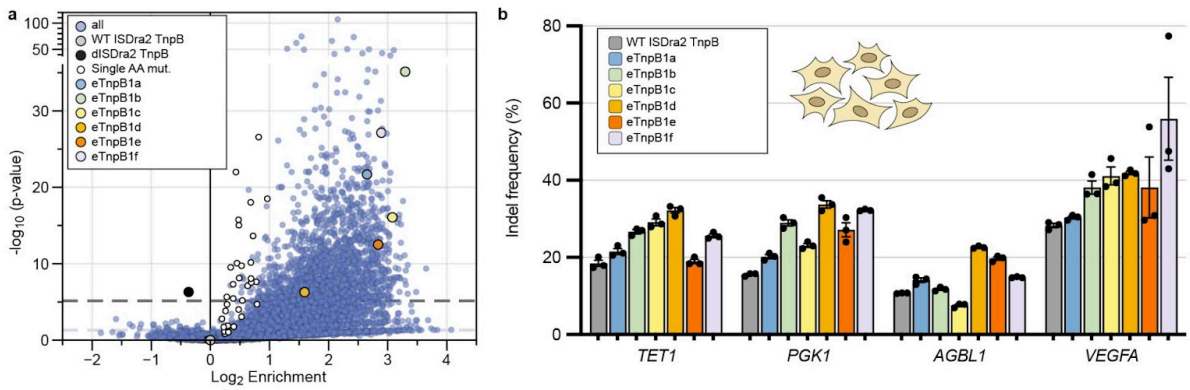
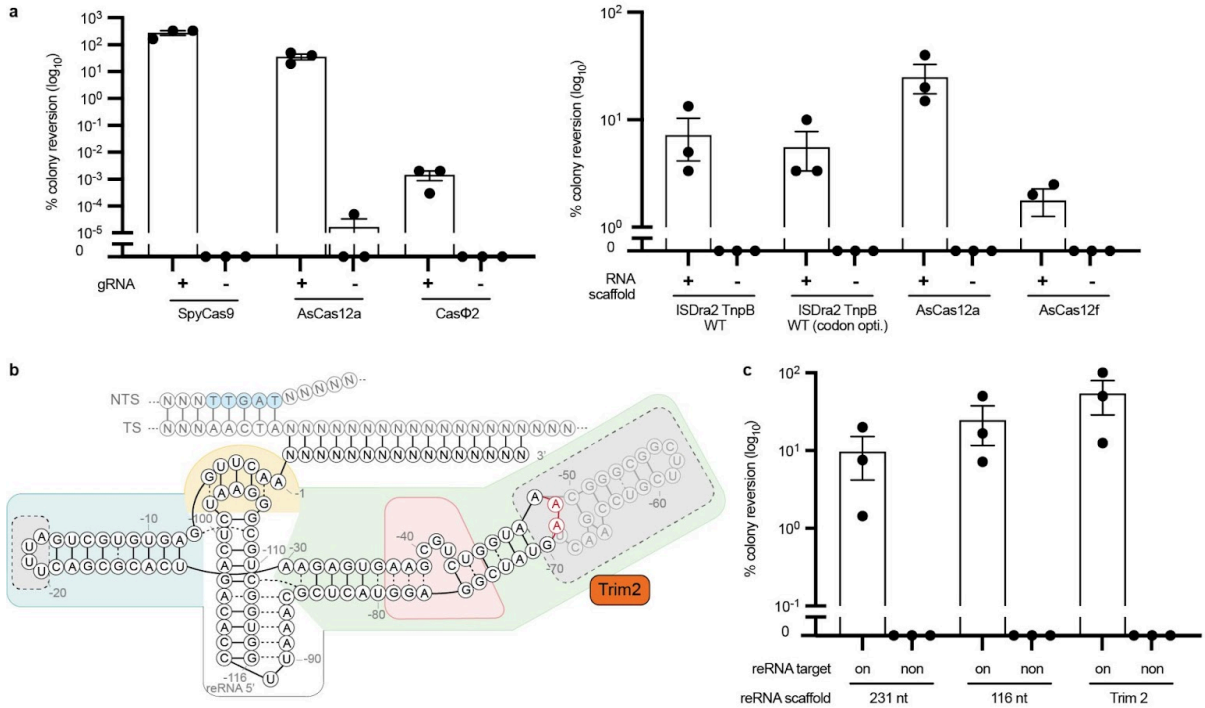


Figure 5. Enhanced TnpB variants engineered by combining high-activity mutations.

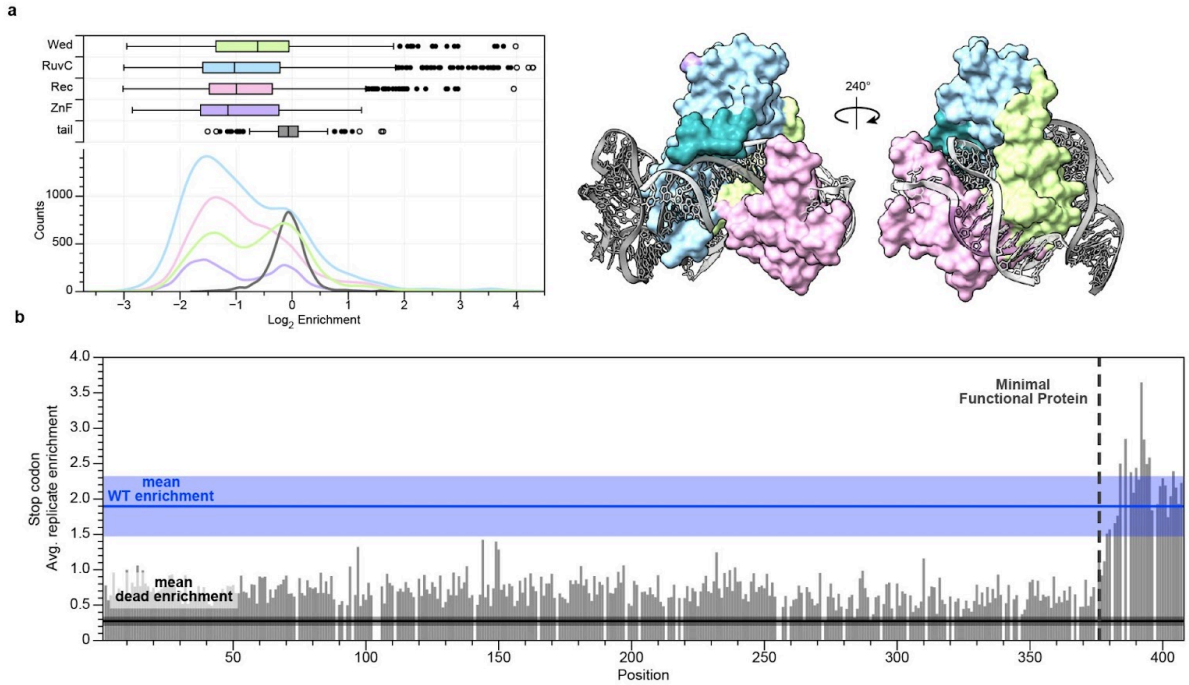
a, Volcano plot depicting combinatorial variant enrichment and statistical significance, after selection in a reporter yeast strain with TnpB target site 1. Enrichment is calculated by taking the average of two biological replicates. Individual barcode enrichments for each variant relative to WT were used to calculate significance with a two-sided Mann-Whitney U-test. Additional characterization of variants is shown in Extended Data Figs. 5-7. **b,** Indel frequency of six combinatorial variants at four genomic loci in HEK293T cells. Plasmids carrying TnpB variants and reRNA were delivered via transfection, and genomic DNA was harvested four days post-transfection and sequenced as described in methods. Data are plotted as the mean and s.e.m. from biological replicates (n=3).

Extended data



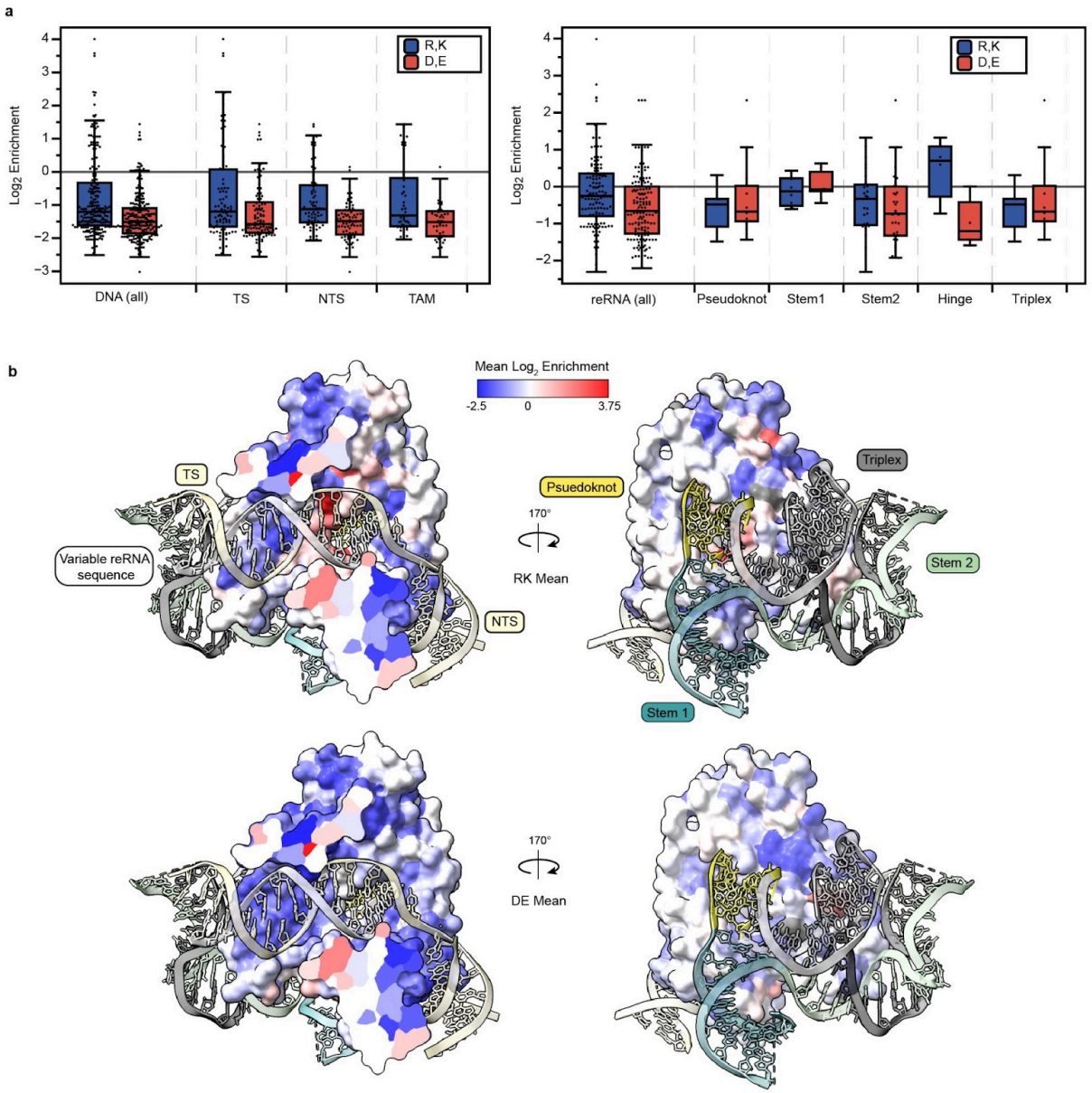
Extended Data Figure 1. *In vivo* yeast cleavage assay captures a range of RNA-guided endonuclease activity across TnpB, Cas9, and Cas12 endonucleases.

a, Assessment of Cas9, Cas12, and ISDra2 TnpB RNA-guided endonuclease activity in yeast cleavage assays, with (+) or without (-) the gRNA or reRNA. ISDra2 TnpB protein was expressed from either the non-codon-optimized open reading frame from *D. radiodurans* R1 (GenBank AE000513.1), or TnpB was codon optimized for high frequency codon usage between *H. sapiens* and *S. cerevisiae.* Data are plotted as the mean and s.e.m. (standard error of mean) from technical triplicate titer plating measurements (n=3). **b,** Schematic representation of the ISDra2 Trim2 reRNA variant (red rArA bases replacing ΔrC^{-50}-rU^{-69}), adapted from Sasnauskas *et al.* 2023[6]. Color scheme corresponds to Fig. 2a. **c,** ISDra2 TnpB endonuclease activity in yeast with various reRNA scaffold lengths, including 231 nts, 116 nts, and the reported Trim2 reRNA variant. TnpB and reRNA were expressed in a yeast strain with an reRNA-complementary target site (on), or in a yeast strain with a non complementary target site (non). Data are plotted as the mean and s.e.m. from technical triplicate titer plating measurements (n=3).



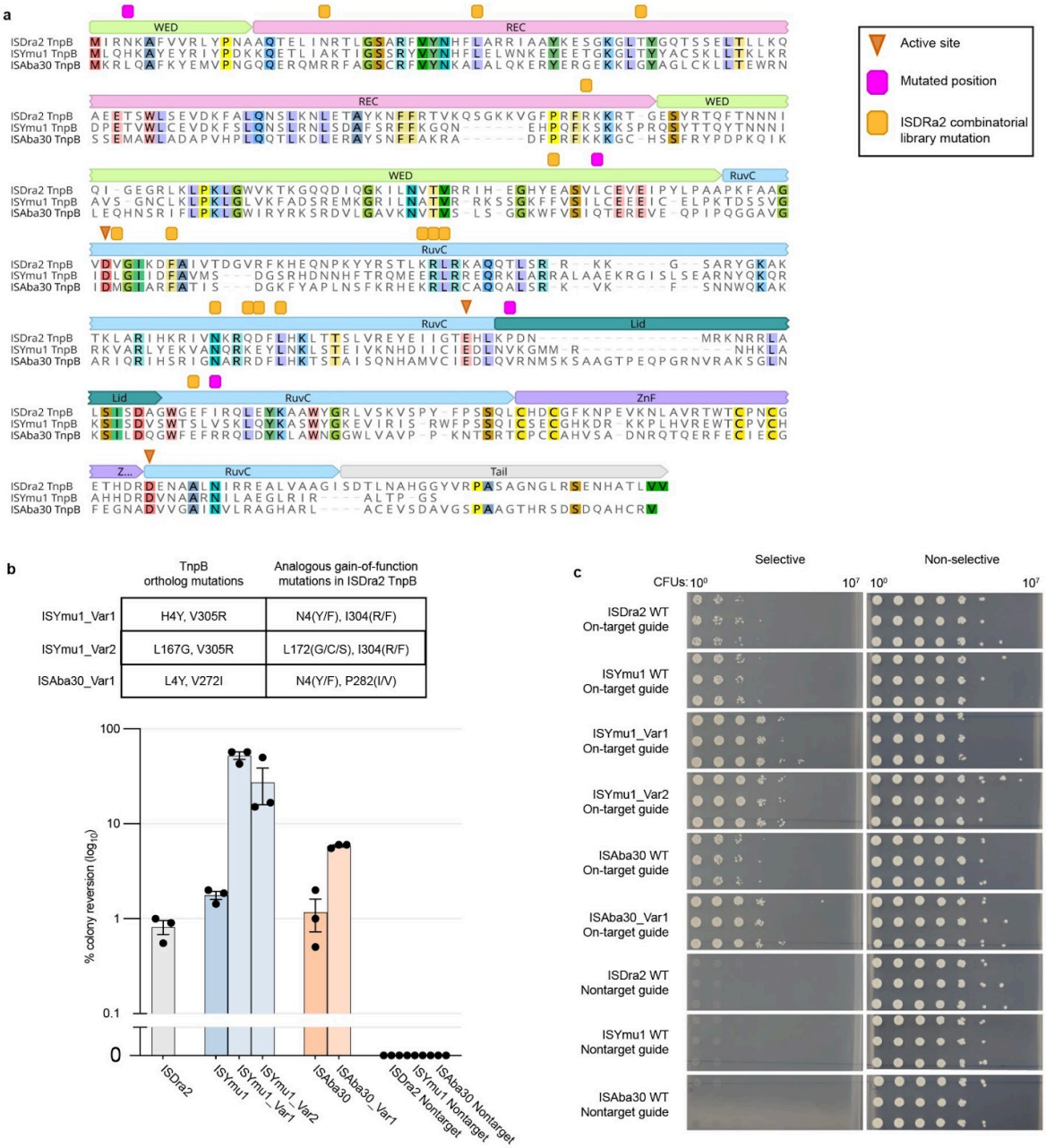
Extended Data Figure 2. Enrichment for single amino acid changes varies by TnpB domains.

a, (Left) Histogram and box plots of the enrichment values for all library mutations, separated by ISDra2 TnpB domain (Right). Many mutations in the RuvC, Rec, and WED domains are enriched above WT, while mutations in the tail domain remain neutral and mutations in the ZnF domain are generally depleted. (Right) ISDra2 domains colored on ISDra2 TnpB ternary structure (PDB ID: 8EXA). Data are plotted as median+IQR. Outliers (Q1-1.5 IQR or Q3+1.5 IQR) drawn as circles and extreme outliers (Q1-3 IQR or Q3+3 IQR) are drawn as open circles **b,** Enrichment scores averaged between both experimental replicates for all stop codon mutations plotted across the length of the protein. Enrichment scores are not normalized to WT. Early stop codons in the tail domain have a neutral effect (enrichment near WT), while stop codons in all other domains are depleted. The dashed line indicates position 376, marking the C-terminus of the minimal active TnpB truncation variant (Δ376–408) previously identified[7].



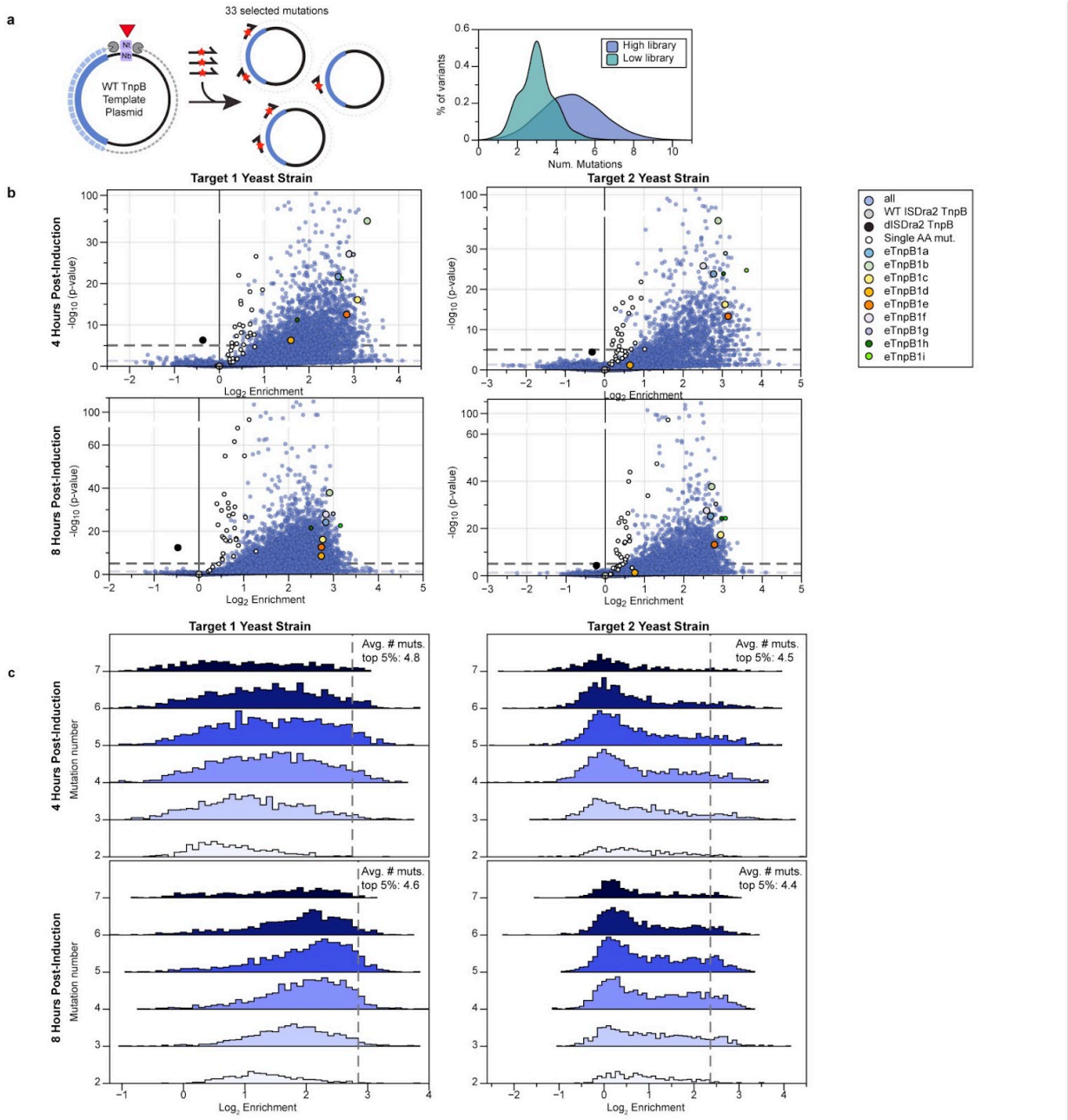
Extended Data Figure 3. Positively charged amino acids are enriched near nucleic acid contacts.

a, Enrichment of positively (R, K) and negatively (D, E) charged amino acid substitutions for residues proximal to nucleic acid. To assess the impact of amino acid substitutions near nucleic acids, we defined proximal residues as those with Cα atoms within 8Å from nucleic acid atoms. This cutoff ensured inclusion of previously identified direct interactions and potential contacts by R/K/D/E mutations. Data are plotted as median + IQR. **b,** The average enrichment for substitutions to positively charged (R, K) or negatively charged (D, E) amino acids was calculated at each position. If the WT residue was already R/K/D/E, its enrichment value was included in the average as 0. Enrichment values were used to color the ISDra2 TnpB cryo-EM ternary structure[6].



Extended Data Figure 4. Activating mutations found for ISDra2 TnpB are transferable to TnpB orthologs.

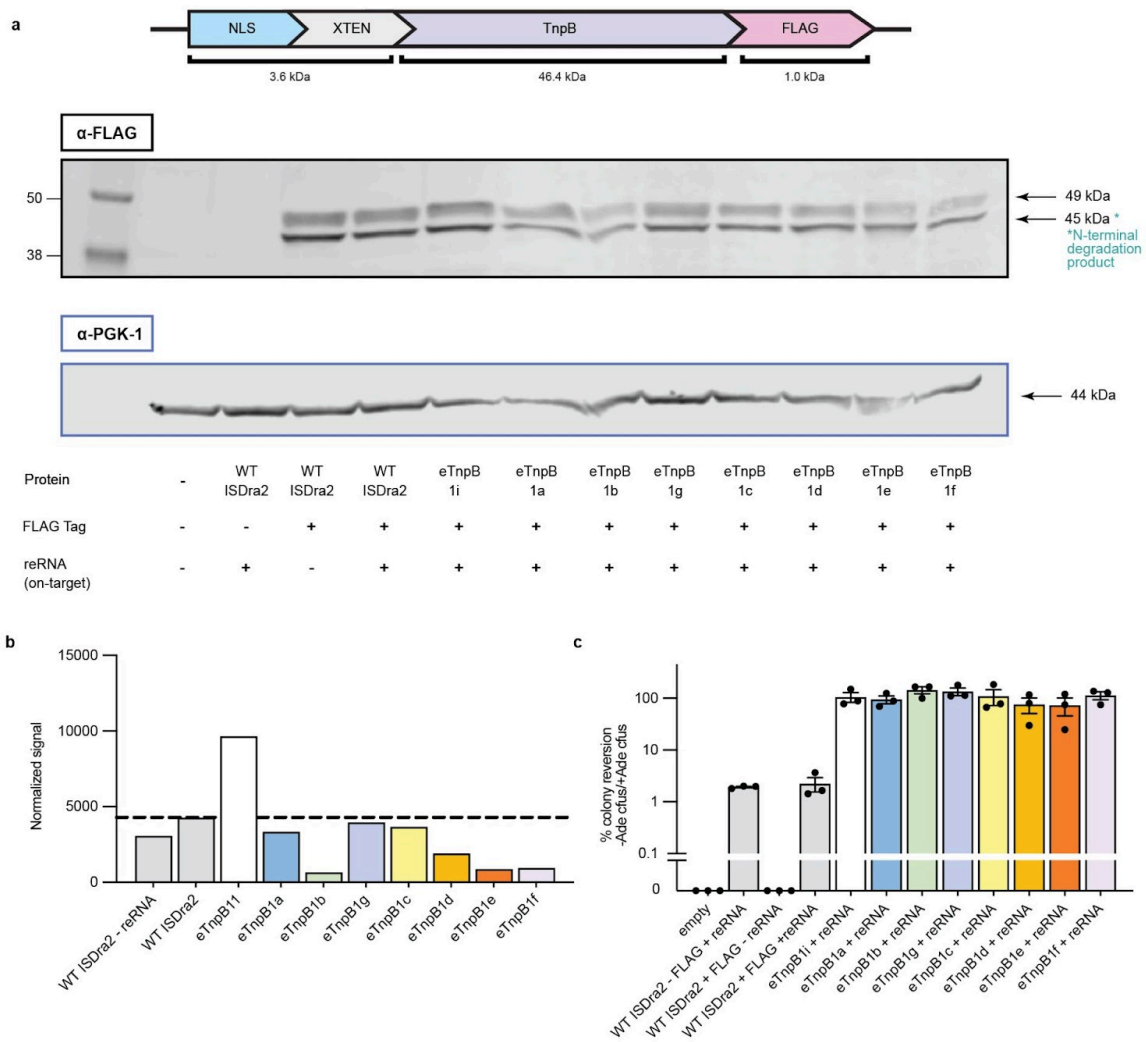
a, Activity of ortholog mutants was assessed by percent colony reversion with the yeast cleavage assay and compared to WT orthologs and negative non complementary reRNA-target controls. Data represent mean ± s.e.m. (n = 3 technical plating replicates 8 hours post-induction). **b,** Titer plates quantified in (a) where each of the protein variants have been expressed in *S. cerevisiae* for 8 hours and titer plated on selective (-adenine) and nonselective (+adenine) media. Technical pipetting replicates from titer plates are shown.



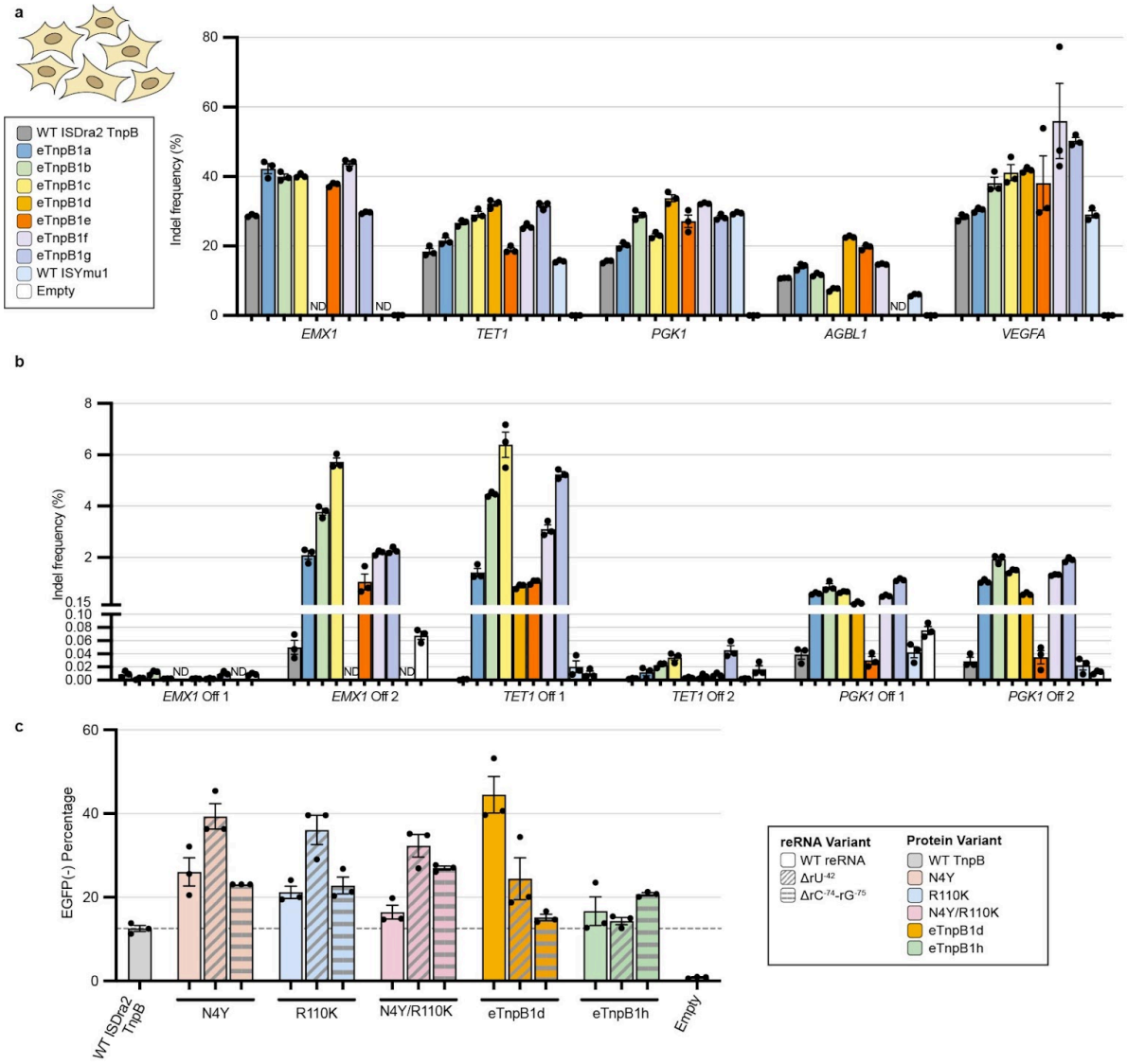
Extended Data Figure 5. Combinatorial library construction, experimental enrichment, and distribution of mutation number.

a, Schematic of library construction using pooled nicking mutagenesis. Plasmid was digested for ssDNA template, and 33 selected amino acid mutations at 19 positions were introduced on ssDNA oligos as described in methods. Two libraries with low and high mutation frequencies were combined, with an average of 3 and 5 mutations, respectively. **b,** Volcano plots of variant enrichment and statistical significance in orthogonal reporter yeast strains with different target sites. Enrichments are shown from 4 hours and 8 hours post-induction. Enrichment is calculated by averaging two biological replicates. Significance was calculated from individual barcode enrichments

per variant relative to wild-type (two-sided Mann-Whitney U-test). **c,** Enrichment scores distributions separated by variant mutation number for experimental conditions in (b).

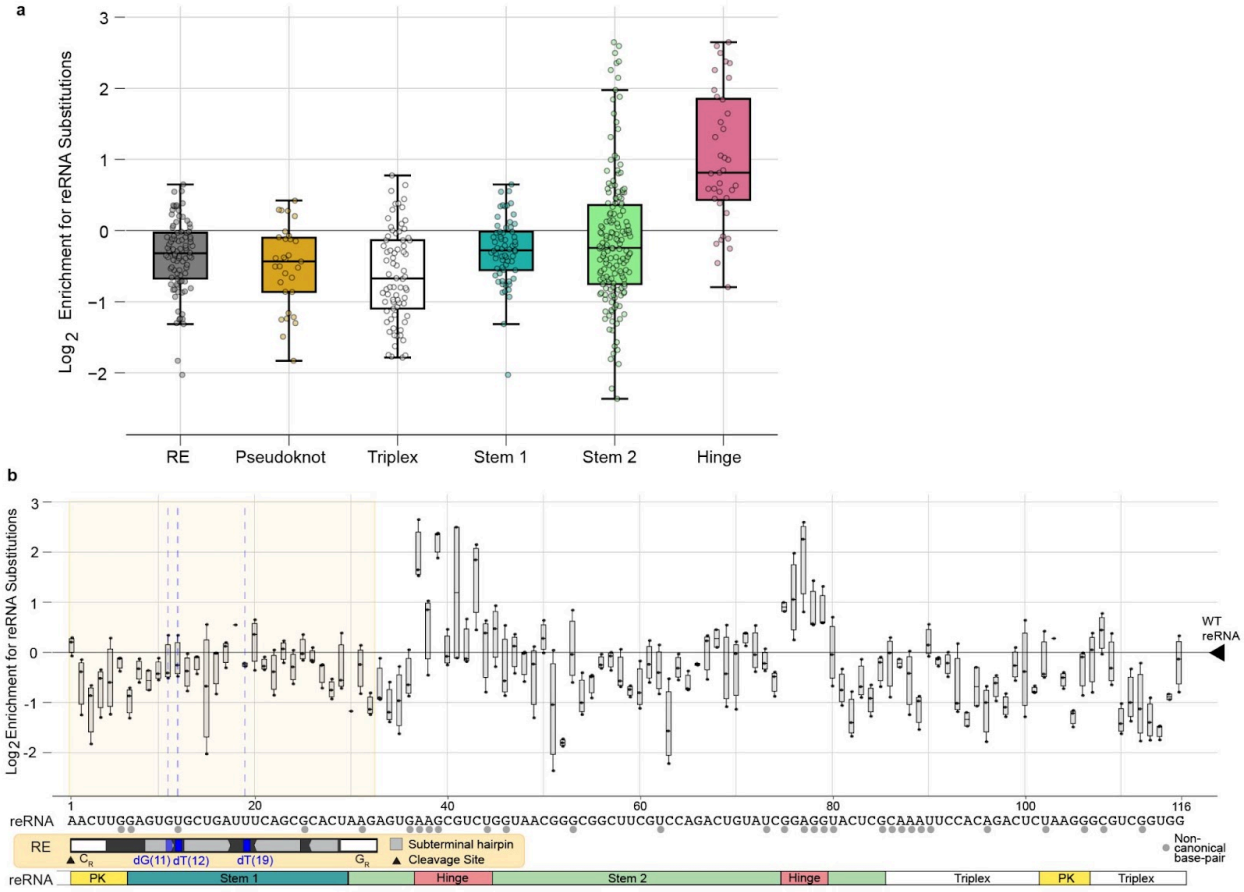


Extended Data Figure 6. Western blots showing that expression levels of enhanced TnpB variants do not increase in *S. cerevisiae*.
a, (Upper) Construct design for expression of ISDra2 TnpB WT protein and variants fused to an NLS and FLAG tag in yeast. (Lower) Western blot from yeast lysate with anti-FLAG antibody, and with an anti-PGK1 antibody as a loading control. **b,** Quantification of signal from Western blot for each variant. **c,** Activity of each variant was assessed by colony reversion in the yeast cleavage assay. Data represent mean ± s.e.m. (n = 3 technical plating replicates).



Extended Data Figure 7. Assessment of combinatorial TnpB variant off and on-target editing, with reRNA mutants in HEK293Ts.

a, Indel frequency of seven combinatorial variants at five genomic loci in HEK293T cells, with WT ISDra2 TnpB, WT ISYmu1 TnpB, and no plasmid controls. Indel frequencies for eTnpB1a-eTnpB1f and WT TnpB at *TET1*, *PGK1*, *AGBL1*, and *VEGFA* are also represented in Fig. 5b. Data are plotted as the mean and s.e.m. from biological replicates (n=3). ND indicates no data. **b,** Indel frequency of WT and combinatorial variant TnBs at off-target sites identified by Cas-OFFinder, with 4–6 mismatches to three on-target sites. Sample order and color scheme match legend in (a). Data are plotted as the mean and s.e.m. from biological replicates (n=3). ND indicates no data. **c,** Pairs of reRNA deletion and ISDra2 TnpB protein mutants were tested with the EGFP KO assay, where EGFP-negative cells were measured by flow cytometry seven days after transfection. Data are presented as the mean ± s.e.m. from biological replicates (n = 3).



Extended Data Figure 8. Deep mutational scanning of reRNA reveals mutational tolerance within the reRNA stem 1-RE overlap.

a, Box and whisker plots showing the distribution of $\log_2$ enrichment for single nucleotide substitutions around the median, normalized to the wild-type reRNA $\log_2$ enrichment. Nucleotide substitutions were grouped by reRNA region. Data are represented as the median + IQR. **b,** Log_2 enrichment of single-nucleotide substitutions at each position in the reRNA, presented as a box-and-whisker plot, with each point representing an individual mutant. The x-axis includes annotations for both the overlapping RE DNA and reRNA sequences. Within the RE, key functional elements are highlighted, including the ssDNA subterminal hairpins recognized by TnpA for transposon excision, as well as the tetranucleotide cleavage (C_R) and guide (G_R) sequences, which form base-pair interactions and direct TnpA cleavage[30,50]. Additionally, positions within the subterminal hairpin important for TnpA binding and strand discrimination are indicated in blue[29].