Title

Engineering in vivo hypermutation and selection systems for observing molecular evolution at scale

Author

Rix, Gordon

Publication Date

2023

UNIVERSITY OF CALIFORNIA,
IRVINE

# Engineering *in vivo* hypermutation and selection systems for observing molecular evolution at scale

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Gordon Rix

Dissertation Committee:
Professor Chang C. Liu, Chair
Professor Jennifer A. Prescher
Professor Andrej Luptak
Associate Professor Han Li
Professor Qing Nie

2023

# Dedication

I dedicate this dissertation to

my father, Kenneth Rix,

who granted me:

a deep appreciation for life’s inherent beauty,

a love of learning,

and infinite opportunity.

# Table of contents

# List of figures

# Acknowledgements

I thank Chang Liu for his mentorship and numerous deep scientific discussions, and for granting me the scientific freedom to pursue my interests. This work would not be possible without his guidance.

I thank all members of the Liu lab, past and present, for your contributions to both this work and my growth as a scientist, and for providing a community that has been invaluable to me.

I thank my family, who have cheered me on every step of the way, for all their love and support.

I thank my ride-or-die Courtney Carlson for sharing this journey (and all future journeys) with me, and for being a limitless source of joy.

I thank Elsevier for permission to include text and figures originally published in *Current Opinions in Chemical Biology*, and Springer Nature for permission to include text and figures originally published in *Nature Communications*.

# Vita

## Gordon Rix

### Education

| | |
|---|---|
| 2017 | B.A. in Cell and Molecular Biology,<br>*University of Rhode Island* |
| 2023 | M.S. in Biological Sciences<br>*University of California, Irvine* |
| 2023 | Ph.D. in Biological Sciences<br>*University of California, Irvine* |

### Research experience

| | |
|---|---|
| 2015-2017 | Graduate Research Assistant<br>*University of Rhode Island* |
| 2017–2023 | Graduate Research Assistant<br>*University of California, Irvine* |

### Teaching Experience

| | |
|---|---|
| 2019 | Graduate Teaching Assistant<br>*University of California, Irvine* |

### Publications

Rix, G., et al. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nature Communications* 11 (1), 5644 (2020)

Rix, G. and Liu, Chang C. Systems for in vivo hypermutation: a quest for scale and depth in directed evolution. *Current Opinion in Chemical Biology* **64**, 20-26 (2021)

Rix, G. et al. In vivo hypermutation and continuous evolution. *Nature Reviews Methods Primers*. **2** (1), 36 (2022)

# Abstract of the dissertation

Engineering *in vivo* hypermutation and selection systems for observing molecular evolution at scale

by

Gordon Rix

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2023

Professor Chang Liu, Chair

*In vivo* hypermutation holds great promise for engineering new biomolecular functions and enabling the study of biomolecular evolution. In this work, we improve the mutagenesis capabilities of our lab's *in vivo* hypermutation system, OrthoRep, and apply it to the evolution of the tryptophan synthase β-subunit TrpB. We first demonstrate that OrthoRep can be used to rapidly evolve this enzyme toward L-tryptophan production in yeast independent of the tryptophan synthase α-subunit. We find that a randomly sampled panel of the resulting enzymes exhibits a broad range of substrate promiscuities, recapitulating the cryptic genetic variation often found in natural protein orthologs.

To facilitate more rapid biomolecular evolution, we next engineered the OrthoRep error prone DNA polymerase toward increased mutation rates and reduced mutation bias, culminating in a set of polymerases that exhibit a mutation rate of $10^{-4}$ substitutions per base per generation, 1-million-fold higher than the native yeast genomic mutation rate. Application of this accelerated *in vivo* hypermutation to TrpB evolution, coupled with a computational pipeline for analysis of the resulting mutation-rich high throughput sequencing datasets, uncovered evidence for structurally distributed interdependence of mutations, as well as indirect evolutionary forces shaping outcomes. Finally, we developed a genetic circuit that enables direct selection for the production of noncanonical amino acids by TrpB. This approach, when combined with *in vivo* hypermutation by OrthoRep and a wealth of engineered TrpBs with altered substrate preferences, has potential in greatly improving the enzymatic noncanonical amino acid production platform of Trp.

# Chapter 1. Introduction to *in vivo* hypermutation

## 1.1 *In vivo* hypermutation

Evolution is the ultimate bioengineer. Yet from the perspective of any individual gene in a modern organism's genome, evolution acts very slowly. This is not just an empirical observation but rather a basic outcome of growing complexity in self-evolving systems. As organisms became more complex over the history of life, genomes gained and relied on more genes. Consequently, the mutation rate per gene had to decrease, because there were more and more things to break. Modern life has thus reached a point where organisms are complex, genomes are large, and the genomic mutation rate must stay low[1–4], too low to rapidly drive the extensive evolution of any particular gene.

To an evolutionary biologist, this may not be a problem, as genomes and many-gene systems can still evolve at a high rate, giving much to observe. But to a protein engineer, this is a major problem. *If the gene encoding a protein must obey the low mutational speed limit of large genomes, how can we watch protein evolution in action? And how can we exploit the tried-and-true power of evolution to make proteins carry out ambitious new functions on laboratory timescales?* In this introductory chapter, we discuss a quickly growing area of research that aims to design and apply genetic systems that selectively mutagenize user-defined genes of interest (GOIs) within living cells[5,6]. Such systems bypass the low mutational speed limits of genomes in order to drive the rapid continuous evolution of GOIs and the proteins they encode simply as cells are passaged under selection.

## 1.2 Motivation

It is useful first to consider why there is a need for *in vivo* mutagenesis and continuous evolution systems when the mature field of directed evolution has already made it possible to evolve GOIs on laboratory timescales. In classical directed evolution, researchers subject one or more GOIs to cycles of *in vitro* diversification (*e.g.* error-prone PCR), transformation of the diversified GOI library into cells, and screening or selection for desired functions. This process can be viewed as a manual bypass of genomic mutational speed limits: one mutates GOIs in a test tube to access the high rates of diversification that large genomes disallow but then transforms GOI variants into cells where they can express and be subjected to functional selection. Yet by manually staging the steps of evolution, classical directed evolution largely forfeits what may be the two most defining features of natural evolution: *scale* and *depth*.

First, scale. Because manual stages of diversification, transformation, and selection are labor-intensive and technically challenging, one can only classically run a few independent evolution experiments at a time, hindering exploration of powerful ideas requiring evolutionary search scale. Such ideas include exploiting spatial structure during protein evolution to escape local optima (*e.g.* the division of a single well-mixed population into many physically separated populations), maximizing diversity of outcomes by running evolution experiments in hundreds if not thousands of independent replicates, obtaining statistical information on evolutionary pathways (*e.g.* understanding drug resistance by mapping rugged fitness landscapes), and evolving a parent protein into families of variants that have different functions (*e.g.* creating versatile collections of antibodies, biosensors, or biosynthetic enzymes).

Second, depth. Classically, one can only take short “walks” on the fitness landscape of a GOI, because a single manual cycle of diversification, transformation, and selection can take days to weeks. Tantalizing ideas requiring evolutionary search depth (*i.e.* long mutational walks) are thus restricted. Such ideas include probing or exploiting the relationship between complex selection histories and adaptation (*e.g.* employing drift to escape local optima during enzyme evolution or using alternating selections to evolve "evolvability")[7] and attaining ambitious protein functions (*e.g.* novel enzyme activities, catalytic activity from *de novo* designed structures, custom function from multi-gene metabolic pathways, or intricate protein-protein or protein-nucleic acid interactions)[8,9] that, by definition, require long mutational pathways to reach. All of these privileges of evolutionary scale and depth are found in abundance in the natural evolution of species – they are responsible for what Darwin described as the "endless forms most beautiful and most wonderful" around us – but have not been captured for laboratory application to engineering GOIs for user-defined functions.

In our view, the key motivation for building genetic systems that achieve targeted hypermutation of GOIs *in vivo* is to transform classical directed evolution into an autonomous and continuous process so that evolutionary search scale and depth become readily available in protein and GOI engineering experiments. This transformation may usher in an era of protein biology and engineering where we can study the process of protein evolution and the underlying sequence-function relationships behind proteins with newfound statistical power, where the evolution of previously difficult protein and GOI functions becomes facile and high-throughput, and where truly novel protein functions fall into the reach of protein evolution and design.

## 1.3 Categorization of *in vivo* hypermutation and continuous evolution systems

The critical task for achieving *in vivo* hypermutation is figuring out how to target rapid mutational accumulation only to GOIs. This targeting is what allows us to bypass the low mutational speed limit of a cell's genome without resorting to the manual staging of diversification and selection in classical directed evolution that restricts scale and depth in evolutionary search. So far, targeting has been achieved through three general architectures that define existing systems: architecture 1 – propagation of viral genomes through non-propagating hosts, architecture 2 – recruitment of mutagenesis machinery to specific DNA loci, and architecture 3 – orthogonal DNA replication. For reasons that will become clear, we call architecture 1 *viral*, architecture 2 *epi-hypermutation*, and architecture 3 *direct hypermutation* (Fig. 1.1). Given our view that the main significance of continuous evolution systems is their admission of scale and depth in protein evolution, we survey these architectures with a focus on their ability to enable evolutionary scale and depth.

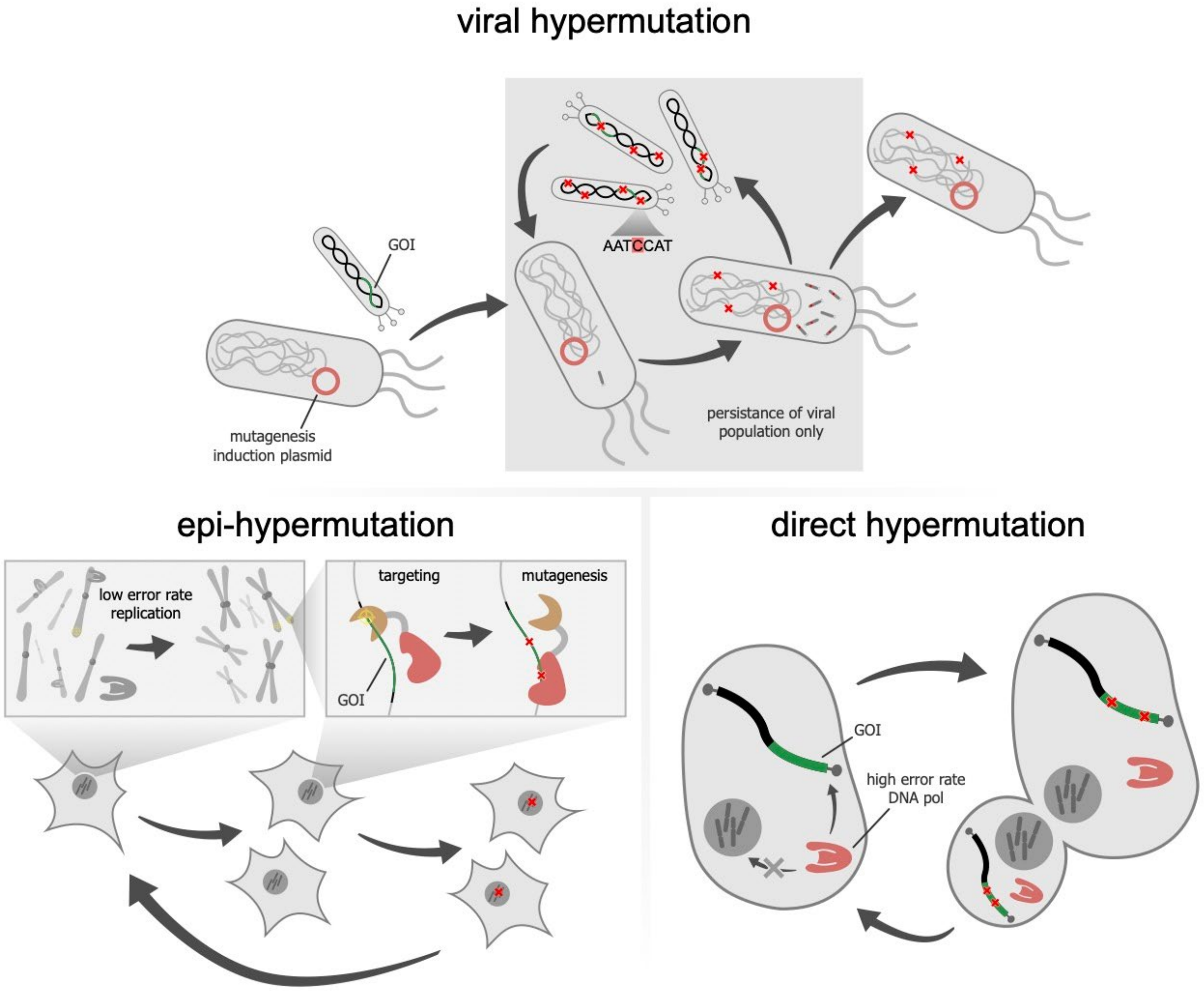


**Figure 1.1. Categories of *in vivo* hypermutation architectures.** Viral hypermutation architectures depend on the small size and therefore high error threshold of viral genomes to enable hypermutation. Epi-hypermutation systems utilize mutagenesis that occurs separately from replication of the target sequence. Direct hypermutation systems utilize mutagenesis that is concurrent with replication of the target sequence

## Viral architectures

In viral architectures for continuous evolution, GOIs are encoded on the genome of a virus where they are induced to hypermutate when propagating through host cells. High rates of mutation can be durably maintained over time because viral genomes are sufficiently small. By coupling the desired activity of the GOI to the ability to make virus, the hypermutating GOI continuously and rapidly evolves as the virus passes through successive hosts.

This is the basic strategy behind phage-assisted continuous evolution (PACE), which is the most mature example of the viral architecture [5,6,10,11]. In PACE, GOIs are encoded on the M13 bacteriophage genome and the phage are propagated through *Escherichia coli* host cells engineered to inducibly mutate at high rates. Although these high mutation rates in the range of $10^{-5}$ to $10^{-4}$ substitutions per base (s.p.b.) are deleterious or lethal to *E. coli*, this is not a problem for the system because the *E. coli* are constantly flowed into and out of a reservoir containing phage; the flow rate is such that only the phage population persists, resulting in the continuous accumulation of mutations only in the phage genome and GOI. To select for desired function from the GOI, the M13 genome is engineered to lack the essential gIII gene whose protein product, pIII, is required for phage packaging and infection. Instead, gIII is encoded in the *E. coli* host where its expression can be made to depend on the desired GOI function. GOI variants with improved function induce higher expression of pIII and more phage descendants, thereby effecting the continuous evolution of the GOI. In this manner, PACE has been used to evolve RNA polymerases [10], biosensors [12], insecticidal proteins [13], base editors [14,15], prime editors [16], metabolic enzymes [17], proteases [18], and more [19–23].

The PACE concept has been extended to mammalian cell hosts as well. Recent work has shown that it is possible to use either an engineered error-prone version of adenovirus or a naturally error-prone RNA virus, sindbis, as vehicles for GOI hypermutation [24–26]. As in PACE, GOIs are encoded on the genome of a virus engineered to lack the essential proteins required for viral production. Instead, the essential proteins are supplied by host mammalian cells where their expression is linked to the GOI's activity. In this manner, adenovirus and sindbis have been used to rapidly evolve transcription factors, GPCRs, conformation-specific anti-GPCR nanobodies by passaging virus through successive cultures of host cells [24,25]. However, others have raised issues with one of these platforms (VEGAS) following their attempts to adopt it,[27] such as an inherent selection for loss of the GOI during propagation, suggesting significant technical hurdles may stand in the way of such systems replicating the success of PACE.

It has become clear, particularly with PACE, that the viral architecture can be highly durable, enabling prolonged continuous or semi-continuous GOI evolution experiments that result in depth of evolutionary search. The scale of evolutionary search using the viral architecture can also be high, but this is an area of ongoing development.[28] The viral architecture is naturally limited by the fact that viruses are not autonomously propagating agents and instead depend on a constant supply of new host cells. This can introduce the requirement for bioreactors, as in PACE where the rate of fresh cell supply demands precise external control to prevent the persistence of host mutations, or technical steps such as centrifugation and filtration to physically separate viral-encoded GOI evolution from host evolution in passaging steps. While others have shown that automation coupled with robotics [28] or millifluidics [29] can reduce these requirements for researcher intervention and enable highly reproducible evolutionary outcomes, such configurations still place large technical demands on researchers. Another challenge in the viral architecture is that the unit of selection is defined by viral production. Although clever genetic circuits can link many GOI functions to viral synthesis,[9,21,30] GOI functions that occur on timescales beyond the viral lifecycle, involve complex host biology, or are meant to change a host cell's (or even a multicellular organism's) physiology are not directly selectable through the viral architecture. To achieve greater evolutionary search scale and expand the types of functions that can be evolved, there is room in the ecosystem of continuous evolution systems for fully *in vivo* architectures.

## Epi-hypermutation architectures

Since cells are autonomously propagating agents, continuous evolution of GOIs fully inside cells supports extensive evolutionary search scale – culturing cells is easy to parallelize into many independent replicates, distinct experiments, or spatially structured populations — as well as depth — culturing cells for many generations is straightforward. Of course, cells have very low mutation rates in order to properly maintain the large information content in their megabase to gigabase genomes, and such low mutation rates (typically $10^{-9}$-$10^{-10}$ s.p.b.) are not enough to rapidly sample diversity at the level of an individual GOI. (A mutation rate of ~$10^{-10}$ s.p.b. would sample only a single mutation every million times a kilobase-sized GOI was replicated.) To exploit the scale and depth of evolution that culturing cells should afford, one must devise systems for targeted hypermutation *in vivo* to speed up GOI evolution.

A popular strategy for building targeted hypermutation systems is to fuse DNA mutating enzymes to site-specific DNA binding proteins. A prominent example is the fusion of nucleotide deaminases to T7 RNA polymerases (RNAPs), where the deaminase induces mutagenesis and the T7 RNAP proteins provide targeting[31–34]. A particular advantage of T7 RNAP is its processivity, which acts to drag the deaminase across an entire GOI or section therein. Likewise, fusion of an error-prone DNA polymerase (DNAP) to a nickase-Cas9 can cause mutagenesis across a stretch of DNA near the Cas9 nick site, as is the basis for the EvolvR system[35,36]. These systems have been used evolve model proteins such as antibiotic resistance genes and cancer drug targets through the simple serial culturing of bacteria, yeast, or mammalian cells under selection[31–36].

We categorize these systems involving the fusion of a DNA mutating enzyme to a DNA binding protein, along with other systems such as TADR, CRISPR-X, ICE, and TaGTEAM[37–40], as epi-hypermutation architectures to emphasize that the GOI targeted for hypermutation is not replicated by the hypermutation system itself. Instead, hypermutation is *epi* to the GOI's propagation, which occurs independently through the high-fidelity replication systems of the host (Fig. 1.1). This feature may have implications on durability and evolutionary search depth achievable. For example, when a GOI is targeted for hypermutation, the very sequence responsible for recruiting hypermutation machinery (*e.g.* T7RNAP promoter) may become corrupted via hypermutation. If the GOI under selection can still propagate and express independently of hypermutation, we speculate that hypermutation may slow over time. Current epi-hypermutation systems also have notable off-target activity, resulting in elevations in the mutation rate of the host genome. This potentially increases the chance of evolution outside the GOI, including selection circuits that link the desired GOI function to cell survival as well as the genes encoding the hypermutation machinery itself. The ongoing application of epi-hypermutation systems will reveal whether these issues affect the goal of enabling scale and depth in the evolution of GOIs.

### Direct hypermutation architectures

Another way to achieve targeted hypermutation of GOIs *in vivo* is to give cells a separate DNA replication system dedicated to the propagation of GOIs. Such an orthogonal DNA replication system would consist of a special DNAP that only replicates a cognate plasmid encoding GOIs. Host DNAPs would replicate the host genome but not the special plasmid, completing orthogonality[41]. If the orthogonal DNAP is then made to be error-prone, the system enforces continuous hypermutation of plasmid-encoded GOIs, driving their rapid evolution *in vivo*.

We categorize orthogonal replication as a direct hypermutation architecture because mutation occurs during replication. This distinction from epi-hypermutation, where mutation of a GOI occurs independently of its replication, may be important for achieving search depth in continuous evolution experiments. For example, if the orthogonal plasmid is hypermutated in a way that prevents its recognition by the orthogonal DNAP, the plasmid isn’t replicated, ensuring that only continuously mutating GOIs persist in an experiment. An additional advantage of having an altogether separate replication system for GOIs is that targeting of hypermutation can be more effectively achieved, for example through spatial separation between the orthogonal and the genomic replication systems.

Early work towards an orthogonal replication system was carried out by Camps *et al.* through the establishment of an error-prone Pol I DNAP that targets ColE1 plasmids[42]. However, since Pol I is also essential for genomic DNA replication, full orthogonality was not achieved. Over the past few years, our lab has developed a fully orthogonal DNA replication (OrthoRep) system by adapting an autonomously replicating cytoplasmic plasmid element found in certain strains of yeast and engineering error-prone variants of the DNAP responsible for replicating the cytoplasmic plasmid.[41] While other groups have recently taken inspiration from this architecutre in establishing an orthogonal replication system in *Bacillus thuringiensis*[43], OrthoRep mutation rates are currently the highest among direct hypermutation systems. Previously developed OrthoRep systems drive the hypermutation of GOIs at a mutation rate of ~$10^{-5}$ s.p.b., while keeping the host *Saccharomyces cerevisiae* genomic mutation rate unchanged at ~$10^{-10}$ s.p.b.[4]. This ~100,000-fold increase in the *in vivo* mutation rate of GOIs has allowed us to evolve several enzymes and proteins with exceptional scale and depth. For example, OrthoRep was used to evolve the malarial drug target, dihydrofolate reductase, in over 100 independent experiments to map mutational pathways leading to drug resistance[4]. More recently,OrthoRep was used to drive the evolution of yeast surface-displayed antibodies in multiple parallel experiments that yielded potent neutralizers of SARS-CoV-2 pseudovirus and conformationally-selective high-affinity nanobodies against a GPCR[44].

OrthoRep is not without its limitations. New versions of the orthogonal DNAP are needed to reach higher mutation rates that can speed up GOI evolution even further. New DNAPs are needed to lower the bias for transition mutations that current error-prone orthogonal DNAPs exhibit. Stronger promoters for hypermutating GOIs on the orthogonal plasmid and new strategies for expression control are desired to improve the range of GOIs and properties that can be evolved [45]. Novel selection strategies that couple arbitrary desired GOI functions to cell fitness with a level of durability that matches the prolonged hypermutation power of OrthoRep are also needed. Additionally, whether OrthoRep can be ported into organisms beyond yeast remains to be seen, in contrast to epi-hypermutations systems, which have already been established in bacteria, yeast, and mammalian cells. We expect that the unique architectural advantages of OrthoRep should provide sufficient motivation for continued development in these directions.

## 1.4 Applications and developments of the OrthoRep system presented in this work

Here, we present the application of the OrthoRep system to enzyme engineering, and the engineering of the OrthoRep system itself to further elevate its mutation rate.

First, we discuss our novel application of OrthoRep to the tryptophan synthase β-subunit from *Thermotoga maritima*, *Tm*TrpB. This promiscuous enzyme, whose various natural orthologs already exhibit a diversity of promiscuities that enable synthesis of L-tryptophan analogs, presented a unique opportunity for OrthoRep. Could we use OrthoRep to evolve a diverse panel of enzymes that recapitulate this valuable feature of natural orthologous TrpBs? Using a simple selection system in which yeast that encode *Tm*TrpB on the OrthoRep plasmid and lack the native tryptophan synthase are challenged to complement tryptophan production, we evolved this enzyme to function in yeast, isolated over 60 of the resulting enzymes, and demonstrated that they exhibit a broad range of promiscuities when tested *in vitro*.

These experiments were very revealing not only of the power of the OrthoRep system, but also of its limitations and unrealized potential. With previously engineered error prone DNAPs, error rates were sufficient to generate 0.1 substitution mutations per 1 kb gene per day on average, leaving 90% of copies of the gene unmutated within this timeframe. Furthermore, mutation rates were highly biased toward transition mutations, which tend to produce more conservative amino acid substitutions. The goal of achieving evolution of new functions at the maximum speed would therefore likely benefit from improved mutation rates. Another enticing application of the OrthoRep system is to generate large datasets of diverse sequences with known functionality to better understand sequence-function relationships. Without enrichment for more mutagenized sequences via strong selection pressure, which can be challenging to maintain over long timescales, generating the level of genetic diversity necessary to produce such datasets would take far too long, *e.g.* 100 days of passaging yeast to achieve 10 additional nucleotide mutations. To address these issues, we engineered the OrthoRep polymerase. Using directed evolution and a dual selection/screening approach, we increased error rates 10-fold and dramatically reduced the mutation bias, yielding an *in vivo* mutation rate 1-million-fold higher than that of the genomic replication machinery, the highest among all reported *in vivo* hypermutation systems. Using the resulting DNAPs and the *Tm*TrpB selection system, we demonstrated that these mutation rates can be sustained over hundreds of generations, enabling not only evolution of *Tm*TrpB function in yeast but producing a large, deeply diversified set of sequences. We developed a high throughput sequencing analysis pipeline and a computational toolkit for more thoroughly profiling the sequence space that had been explored. Using this framework, we studied the resulting evolutionary outcomes and found examples of strong interdependence of mutations to distant residues, likely related to allosteric coordination of catalysis, and uncovered a selection pressure for decreased isoelectric point.

Finally, we developed a growth-based selection for expanding the substrate scope of TrpB. Where our prior work depended on selection only for the native tryptophan synthesis activity, here we describe a strategy that specifically enriches for the production of noncanonical amino acid analogs using a synthetic genetic circuit. Using a recently engineered *Tm*TrpB, deemed *Tm*TyrS6, as a starting point, we put this selection to the test. Using our library of diversified Trp-producing *Tm*TrpBs as a source for generally activating mutations, and coupling this library to our genetic circuit, we isolated *Tm*TyrS variants that could produce 3-iodo-L-tyrosine *in vivo*.

We believe this work, strengthened by the advantages afforded by the direct hypermutation architecture of OrthoRep, marks a significant advancement in the field of *in vivo* hypermutation. While it highlights some key areas where further development of the technology is needed, it also enables many exciting directions for a wide range of applications. We conclude with a discussion of these matters.

## 1.5 References

1. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 7160–7164 (1991).
2. Biebricher, C. K. & Eigen, M. What is a quasispecies? *Curr. Top. Microbiol. Immunol.* **299**, 1–31 (2006).
3. Herr, A. J. *et al.* Mutator suppression and escape from replication error-induced extinction in yeast. *PLoS Genet.* **7**, (2011).
4. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell.* **175**, 1946-1957.e13 (2018).
5. Simon, A. J., d'Oelsnitz, S. & Ellington, A. D. Synthetic evolution. *Nat. Biotechnol.* (2019)

doi:10.1038/s41587-019-0157-4.

6. Morrison, M. S., Podracky, C. J. & Liu, D. R. The developing toolkit of continuous directed evolution. *Nat. Chem. Biol.* **16**, 610–619 (2020).

7. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).

8. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nat. Catal.* **3**, 203–213 (2020).

9. Zinkus-Boltz, J., Devalk, C. & Dickinson, B. C. A Phage-Assisted Continuous Selection Approach for Deep Mutational Scanning of Protein-Protein Interactions. *ACS Chem. Biol.* **14**, 2757–2767 (2019).

10. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature.* **472**, 499–503 (2011).

11. Miller, S. M., Wang, T. & Liu, D. R. Phage-assisted continuous and non-continuous evolution. *Nat. Protoc.* **15**, 4101–4127 (2020).

12. Pu, J., Zinkus-Boltz, J. & Dickinson, B. C. Evolution of a split RNA polymerase as a versatile biosensor platform. *Nat. Chem. Biol.* **13**, 432–438 (2017).

13. Badran, A. H. *et al.* Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature.* **533**, 58–63 (2016).

14. Thuronyi, B. W. *et al.* Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* **37**, (2019).

15. Richter, M. F. *et al.* Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).

16. Doman, J. L. *et al.* Phage-assisted evolution and protein engineering yield compact, efficient prime editors. *Cell.* **186**, 3983–4002 (2023).

17. Roth, T. B., Woolston, B. M., Stephanopoulos, G. & Liu, D. R. Phage-Assisted Evolution of Bacillus methanolicus Methanol Dehydrogenase 2. *ACS Synth. Biol.* **8**, 796–806

(2019).

18. Packer, M. S., Rees, H. A. & Liu, D. R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat. Commun.* **8**, 956 (2017).

19. Brödel, A. K. *et al.* Accelerated evolution of a minimal 63-amino acid dual transcription factor. *Sci. Adv.* **6**, 1–10 (2020).

20. Ye, X. *et al.* Using phage-assisted continuous evolution (PACE) to evolve human PD1. *Exp. Cell Res.* **396**, 112244 (2020).

21. Bryson, D. I. *et al.* Continuous directed evolution of aminoacyl-tRNA synthetases to alter amino acid specificity and enhance activity. *Nat. Chem. Biol.* (2017) doi:10.1038/nchembio.2474.

22. DeBenedictis, E. A., Chory, E. J., Gretton, D., Wang, B. & Esvelt, K. A high-throughput platform for feedback-controlled directed evolution. *bioRxiv* 2020.04.01.021022 (2020) doi:10.1101/2020.04.01.021022.

23. Schmidheini, L. *et al.* Continuous directed evolution of a compact Cj Cas9 variant with broad PAM compatibility. (2023) doi:10.1038/s41589-023-01427-x.

24. Berman, C. M. *et al.* An Adaptable Platform for Directed Evolution in Human Cells. *J. Am. Chem. Soc.* **140**, 18093–18103 (2018).

25. English, J. G. *et al.* VEGAS as a Platform for Facile Directed Evolution in Mammalian Cells. *Cell.* **178**, 748-761.e17 (2019).

26. Jewel, D., Pham, Q. & Chatterjee, A. Virus-assisted directed evolution of biomolecules. *Curr. Opin. Chem. Biol.* **76**, 102375 (2023).

27. Denes, C. E. *et al.* The VEGAS Platform Is Unsuitable for Mammalian Directed Evolution. *ACS Synth. Biol.* **11**, 3544–3549 (2022).

28. DeBenedictis, E. A. *et al.* Systematic molecular evolution enables robust biomolecule discovery. *Nat. Methods.* **19**, 55–64 (2022).

29. Huang, T. P. *et al.* High-throughput continuous evolution of compact Cas9 variants

targeting single-nucleotide-pyrimidine PAMs. *Nat. Biotechnol.* **41**, 96–107 (2023).

30. Wang, T., Badran, A. H., Huang, T. P. & Liu, D. R. Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.* (2018) doi:10.1038/s41589-018-0121-5.

31. Moore, C. L., Papa, L. J. & Shoulders, M. D. A Processive Protein Chimera Introduces Mutations across Defined DNA Regions in Vivo. *J. Am. Chem. Soc.* **140**, 11560–11564 (2018).

32. Chen, H. *et al.* Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. *Nat. Biotechnol.* **38**, 165–168 (2020).

33. Álvarez, B., Mencía, M., de Lorenzo, V. & Fernández, L. Á. In vivo diversification of target genomic sites using processive base deaminase fusions blocked by dCas9. *Nat. Commun.* **11**, 6436 (2020).

34. Park, H. & Kim, S. Gene-specific mutagenesis enables rapid continuous evolution of enzymes in vivo. *Nucleic Acids Res.* **49**, e32–e32 (2021).

35. Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature.* **560**, 248–252 (2018).

36. Tou, C. J., Schaffer, D. V. & Dueber, J. E. Targeted Diversification in the S. cerevisiae Genome with CRISPR-Guided DNA Polymerase i. *ACS Synth. Biol.* **9**, 1911–1916 (2020).

37. Yi, X., Khey, J., Kazlauskas, R. J. & Travisano, M. Plasmid hypermutation using a targeted artificial DNA replisome. *Sci. Adv.* **7**, (2021).

38. Hess, G. T. *et al.* Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods.* **13**, 1036–1042 (2016).

39. Crook, N. *et al.* In vivo continuous evolution of genes and pathways in yeast. *Nat. Commun.* **7**, (2016).

40. Finney-Manchester, S. P. & Maheshri, N. Harnessing mutagenic homologous

recombination for targeted mutagenesis in vivo by TaGTEAM. *Nucleic Acids Res.* **41**, 1–10 (2013).

41. Ravikumar, A., Arrieta, A. & Liu, C. C. Supplementary Information for: An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

42. Camps, M., Naukkarinen, J., Johnson, B. P. & Loeb, L. A. Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9727–9732 (2003).

43. Tian, R. *et al.* Engineered bacterial orthogonal DNA replication system for continuous evolution. *Nat. Chem. Biol.* (2023) doi:10.1038/s41589-023-01387-2.

44. Wellner, A. *et al.* Rapid generation of potent antibodies by autonomous hypermutation in yeast. *Nat. Chem. Biol.* **17**, 1057–1064 (2021).

45. Zhong, Z., Ravikumar, A. & Liu, C. C. Tunable Expression Systems for Orthogonal DNA Replication. *ACS Synth. Biol.* **7**, 2930–2934 (2018).

# Chapter 2. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities

## 2.1 Abstract

Enzyme orthologs sharing identical primary functions can have different promiscuous activities. While it is possible to mine this natural diversity to obtain useful biocatalysts, generating comparably rich ortholog diversity is difficult, as it is the product of deep evolutionary processes occurring in a multitude of separate species and populations. Here, we take a first step in recapitulating the depth and scale of natural ortholog evolution on laboratory timescales. Using a continuous directed evolution platform called OrthoRep, we rapidly evolved the *Thermotoga maritima* tryptophan synthase β-subunit (*Tm*TrpB) through multi-mutation pathways in many independent replicates, selecting only on *Tm*TrpB's primary activity of synthesizing L-tryptophan from indole and L-serine. We find that the resulting sequence-diverse *Tm*TrpB variants span a range of substrate profiles useful in industrial biocatalysis and suggest that the depth and scale of evolution that OrthoRep affords will be generally valuable in enzyme engineering and the evolution of new biomolecular functions.

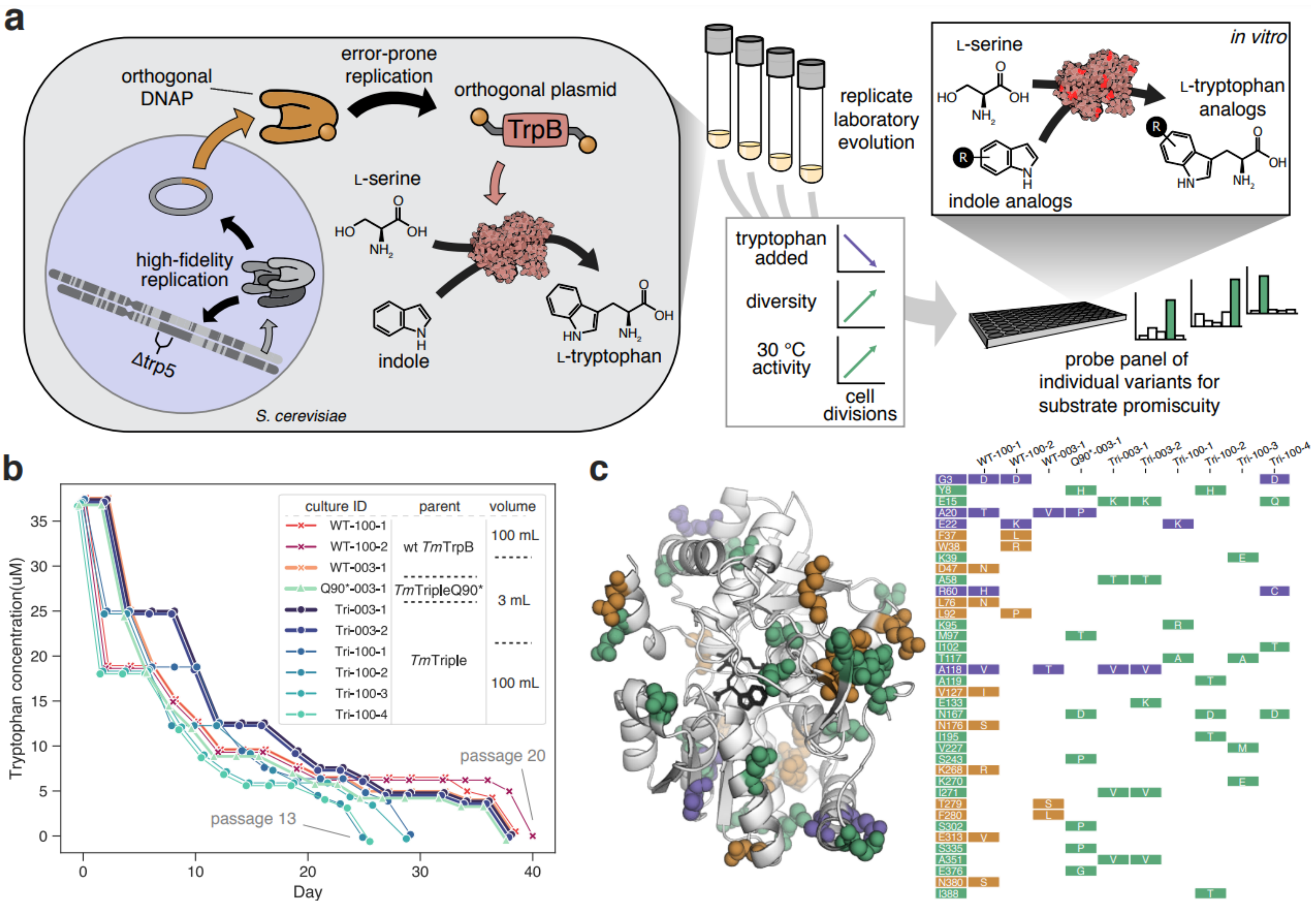
## 2.2 Introduction

Natural enzymes typically have many orthologs. While the primary activity of orthologous enzymes is largely the same,[1] promiscuous functions not under selective pressure can vary widely.[2,3] Such variation may be attributed to the deep and distinct evolutionary histories shaping each ortholog, including long periods of neutral drift, recalibration of primary activity, and adaptation to new host environments such as temperature. These rich histories act to produce extensive genetic diversity, which underpins different promiscuity profiles.[2]

Diversity in promiscuous functions across orthologs is of both fundamental and practical importance. An enzyme's reserve of promiscuous activities dictates what secondary reactions, environmental changes, or niches the enzyme can accommodate.[4,5] Diversity in promiscuous activities therefore contributes to the basic robustness of life and adaptation. An enzyme's reserve of promiscuous activities can also be mined in the application of enzymes for biocatalysis.[6,7] Ortholog diversity therefore expands the range of reactions at the disposal of enzyme engineers, supporting the growing role of "green" enzymatic processes in the chemical and pharmaceutical industries.[8–10]

Inspired by the remarkable ability of enzyme orthologs to encompass promiscuous activities, we asked whether we could extend the substrate scope of useful enzymes by evolving multiple versions of an enzyme in the laboratory, selecting only for its primary function. Although this idea has been explored before using classical directed evolution approaches, most notably through the generation of cryptic genetic variation with neutral drift libraries,[11–14] we recognized that our recently developed continuous evolution system, OrthoRep, may be considerably better poised for this challenge.[15,16] Classical directed evolution mimics evolution through an iterative procedure that involves diversifying a gene of interest (GOI) *in vitro* (*e.g.*, through error-prone PCR), transforming the resulting GOI library into cells, and selecting or screening for desired activities, where each cycle of this procedure represents one step in an evolutionary search.[17] However, since each cycle is manually staged, classical directed evolution does not readily admit depth and scale during exploration of functional sequence space — it is difficult to carry out many iterations to mimic lengthy evolutionary searches (depth), let alone do so in many independent experiments (scale). Yet evolutionary depth and scale are precisely the two characteristics responsible for ortholog diversity in nature. Natural orthologs have diversified from their ancestral parent over great evolutionary timescales, allowing for the traversal of long mutational pathways shaped by complex selection histories (depth). Natural orthologs are also the result of numerous independent evolutionary lineages, since spatially separated species and populations are free to take divergent mutational paths and experience different environments (scale). Systems that better mimic the depth and scale of natural enzyme evolution, but on laboratory timescales, are thus needed for the effective generation of enzyme variants that begin to approach the genetic and promiscuity profile diversity of orthologs.

OrthoRep is such a system. In OrthoRep, an orthogonal error-prone DNA polymerase durably hypermutates an orthogonal plasmid (p1) without raising the mutation rate of the host *Saccharomyces cerevisiae* genome.[16] Thus, GOIs encoded on p1 rapidly evolve when cells are simply passaged under selection. By reducing the manual stages of classical directed evolution down to a continuous process where cycles of diversification and selection occur autonomously *in vivo*, OrthoRep readily accesses depth and scale in evolutionary search.[16,18] Here, we apply OrthoRep to the evolution of the *Thermotoga maritima* tryptophan synthase β-subunit (*Tm*TrpB) in multiple independent continuous evolution experiments, each carried out for at least 100 generations. While we only pressured *Tm*TrpB to improve its primary activity of coupling indole and serine to produce tryptophan, the large number of independent evolution experiments we ran (scale) and the high degree of adaptation in each experiment (depth) resulted in a panel of variants encompassing expanded promiscuous activity with indole analogs. In addition to the immediate value of these newly evolved *Tm*TrpBs in the synthesis of tryptophan analogs, our study offers a new template for enzyme engineering where evolutionary depth and scale is leveraged on laboratory timescales to generate effective variant collections covering broad substrate scope.



| | WT-100-1 | WT-100-2 | WT-003-1 | Q90*-003-1 | Tri-003-1 | Tri-003-2 | Tri-100-1 | Tri-100-2 | Tri-100-3 | Tri-100-4 |
|---|---|---|---|---|---|---|---|---|---|---|
| G3 | D | D | | | | | | | | D |
| Y8 | | | | H | | | | H | | |
| E15 | | | | | K | K | | | | Q |
| A20 | T | | V | P | | | | | | |
| E22 | | K | | | | | K | | | |
| F37 | | L | | | | | | | | |
| W38 | | R | | | | | | | | |
| K39 | | | | | | | | | E | |
| D47 | N | | | | | | | | | |
| A58 | | | | | T | T | | | | |
| R60 | H | | | | | | | | | C |
| L76 | N | | | | | | | | | |
| L92 | | P | | | | | | | | |
| K95 | | | | | | | R | | | |
| M97 | | | | T | | | | | | |
| I102 | | | | | | | | | | T |
| T117 | | | | | | | A | | A | |
| A118 | V | | T | | V | V | | | | |
| A119 | | | | | | | | T | | |
| V127 | I | | | | | | | | | |
| E133 | | | | | | K | | | | |
| N167 | | | | D | | | | D | | D |
| N176 | S | | | | | | | | | |
| I195 | | | | | | | | T | | |
| V227 | | | | | | | | | M | |
| S243 | | | | P | | | | | | |
| K268 | R | | | | | | | | | |
| K270 | | | | | | | | | E | |
| I271 | | | | | V | V | | | | |
| T279 | | | S | | | | | | | |
| F280 | | | L | | | | | | | |
| S302 | | | | P | | | | | | |
| E313 | V | | | | | | | | | |
| S335 | | | | P | | | | | | |
| A351 | | | | | V | V | | | | |
| E376 | | | | G | | | | | | |
| N380 | S | | | | | | | | | |
| I388 | | | | | | | | T | | |

**Fig. 2.1. OrthoRep-mediated continuous *in vivo* evolution of *Tm*TrpB to generate many diverse, functional variants. a** Pipeline for the use of OrthoRep continuous directed evolution to generate many diverse, functional *Tm*TrpB sequences. *Tm*TrpB variants are first evolved in replicate for Trp production in yeast. OrthoRep enables replicate evolution through error-prone replication of an orthogonal plasmid by an orthogonal polymerase, maintaining low error rates in genome replication. By encoding a *Tm*TrpB variant on this plasmid in a tryptophan synthase (TRP5) deletion mutant, *Tm*TrpB may be both continuously diversified and selected for through gradual reduction in Trp supplied in the growth medium. Evolved populations containing many diverse, functional individuals may then be randomly sampled and tested for activity with indole analogs. *Tm*TrpB illustration generated using Illustrate.[40] **b** Selection trajectories for ten replicate cultures that evolved sufficient *Tm*TrpB activity to support cell growth without supplemented Trp. Each point represents a single 1:100 dilution (passage) into fresh indole-supplemented growth medium. Trp concentration of fresh media was reduced when high saturation was achieved in the previous passage. Plots are slightly offset from true values to allow for visibility of all selection trajectories. *Tm*3C and *Tm*3D are plotted as one line as their trajectories were identical. **c** *Tm*TrpB homology model and table depicting consensus mutations of the ten cultures shown in panel **b**. Mutations are colored by their appearance in populations evolved from wt *Tm*TrpB (orange), *Tm*Triple or *Tm*TripleQ90* (green), or both (purple).

## 2.3 Results

### Establishing a selection system for the evolution of *Tm*TrpB variants

To evolve *Tm*TrpB variants using OrthoRep, we first needed to develop a selection where yeast would rely on *Tm*TrpB's primary enzymatic activity for growth. *Tm*TrpB catalyzes the PLP-dependent coupling of L-serine and indole to generate L-tryptophan (Trp) in the presence of the tryptophan synthase α-subunit, *Tm*TrpA.[19] In *T. maritima* and all other organisms that contain a heterodimeric tryptophan synthase complex, TrpA produces the indole substrate that TrpB uses and the absence of TrpA significantly attenuates the activity of TrpB through loss of allosteric activation.[19,20] TRP5 is the *S. cerevisiae* homolog of this heterodimeric enzyme complex, carrying out both TrpA and TrpB reactions and producing Trp for the cell. We reasoned that by deleting the *TRP5* gene and forcing *S. cerevisiae* to rely on *Tm*TrpB instead, cells would be pressured to evolve high stand-alone *Tm*TrpB activity in order to produce the essential amino acid Trp in indole-supplemented media (Fig. 2.1a). This selection pressure would also include thermoadaptation, as yeast grow at mesophilic temperatures in contrast to the thermophilic source of *Tm*TrpB. Therefore, the selection on *Tm*TrpB's primary activity would be multidimensional — stand-alone function, temperature, and neutral drift implemented when desired — and could result in complex evolutionary pathways that serve our goal of maximizing functional variant diversity across replicate evolution experiments. In addition, the multidimensional selection also serves practical goals as stand-alone activity is useful in biosynthetic applications (enzyme complexes are difficult to express and use *in vitro*) and activity at mesophilic temperatures is more compatible with heat-labile substrates, industrial processes where heating costs can compound, or *in vivo* applications in model mesophilic hosts (*e.g. S. cerevisiae* or *Escherichia coli*).

To test this selection, we turned to a positive control *Tm*TrpB variant called *Tm*Triple. This variant was previously engineered to enable stand-alone activity, free from dependence on allosteric activation by TrpA, through a minimal set of three mutations.[7] We found that *Tm*Triple rescued TRP5 function in a *Δtrp5* strain in an indole-dependent manner, validating our selection (Fig. S2.1). Notably, *Tm*Triple, along with other TrpB variants tested, only supported complementation when expressed from a high-strength promoter (Fig. S2.1). This highlighted the opportunity for substantial adaptation and drift even in evolution experiments that start from already engineered *Tm*TrpB variants.

## Continuous evolution of TmTrpB with depth and scale

We encoded wild-type (wt) *Tm*TrpB, *Tm*Triple, as well as a nonsense mutant of *Tm*Triple, *Tm*TripleQ90*, onto OrthoRep's p1 plasmid, which is replicated by a highly error-prone orthogonal DNA polymerase. *Tm*TripleQ90* was included because reversion of the stop codon at position 90 in *Tm*TripleQ90* would act as an early indication that adaptation was occurring, giving us confidence to continue passaging our evolution experiments for several weeks to maximize evolutionary search depth. In all three OrthoRep Δ*trp5* strains, the initial *Tm*TrpB sequences enabled only minimal indole-dependent complementation (Fig. S2.1). This was expected for wt

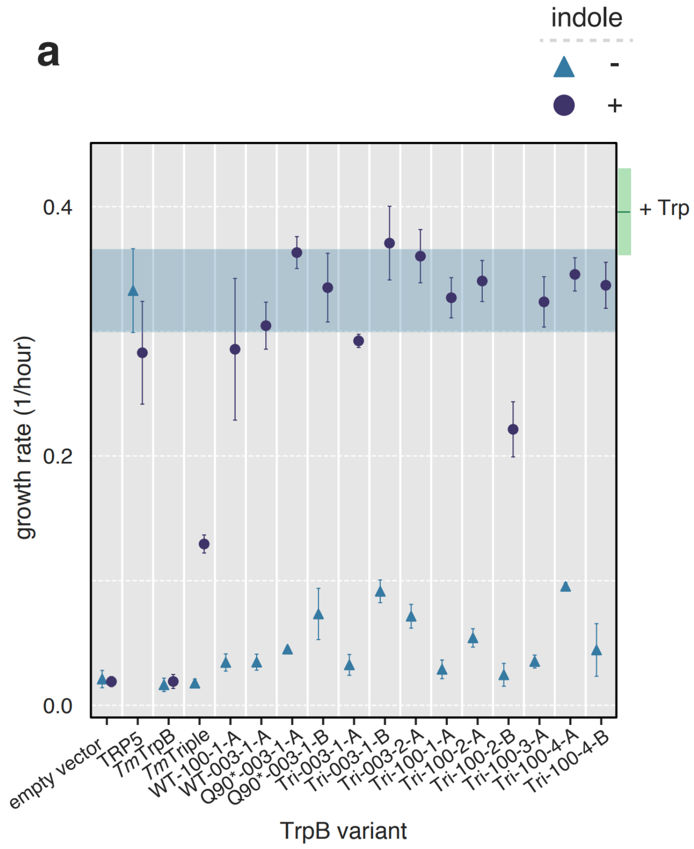
*Tm*TrpB, which has low stand-alone enzymatic activity and *Tm*TripleQ90*, which has a premature stop codon; and was unsurprising for *Tm*Triple, since *Tm*Triple displayed indole-dependent complementation only when artificially overexpressed (Fig. S2.1).

To continuously evolve *Tm*TrpB, we passaged cells encoding wt *Tm*TrpB, *Tm*Triple, or *Tm*TripleQ90* on OrthoRep in the presence of 100 µM indole while reducing the amount of Trp in the medium over time. In total, six 100 mL and twenty 3 mL cultures were passaged, each representing a single independent evolutionary trajectory. Passages were carried out as 1:100 dilutions where Trp concentrations were decreased in the (N+1)$^{th}$ passage if cells grew quickly in the N$^{th}$ passage, until Trp was fully omitted. All six of the 100 mL cultures, and four of the twenty 3 mL cultures fully adapted and were capable of robust growth in indole-supplemented media lacking Trp after 90–130 generations (13–20 passages) (Fig. 2.1b, Table S2.1). Populations that did not achieve growth in the absence of Trp still adapted, but stopped improving at ~5 µM supplemented Trp, suggesting a suboptimal local fitness maximum that is more easy to escape through the greater sequence diversity represented in larger populations. This could explain the different success rates in reaching full adaptation between the 3 mL and 100 mL populations. Cultures that did adapt fully were passaged for an additional ~40 generations without increasing selection stringency to allow for accumulation of further diversity through neutral drift.

For each of the 10 fully adapted populations, we PCR-amplified and bulk-sequenced the *Tm*TrpB alleles on the p1 plasmid. Mutations relative to the parent *Tm*TrpB variant detected at >50% frequency in each population were deemed consensus mutations for that population, with the exception of reversion of the stop codon in populations evolving *Tm*TripleQ90*. This stop codon reversion occurred at 100% frequency in the relevant populations and was not counted in any subsequent analyses due to its triviality. An average of 5.6 (± 2.3 s.d.) and a range of 3–11 consensus amino acid changes per population were observed (Fig. 2.1c, Table S2.2). Some of these mutations occurred at residues previously identified as relevant in conformational dynamics (*e.g.*, N167D and S302P).[20–22] Most mutations observed, however, have not been previously identified in laboratory engineering experiments, suggesting that even the consensus of these populations explored new regions of *Tm*TrpB's fitness landscape, doing so with diversity across replicates (Fig. 2.1c) that might translate to diversity in promiscuous activities across evolved variants.

### Evolved TmTrpB variants improve Trp production in vivo and contain cryptic genetic variation

To ensure that evolved *Tm*TrpB variants, and not potential host genomic mutations, were primarily responsible for each population's adaptation, we cloned individual *Tm*TrpBs into a standard low copy yeast nuclear plasmid under a promoter that approximates expression from p1,[23,24] transformed the variants into a fresh Δ*trp5* strain, and tested for their ability to support indole-dependent growth in the absence of Trp (Figs. 2.2 and S2.2). Sixteen *Tm*TrpB mutants were tested, representing one or two individual variants from each of the ten fully adapted populations. We found that 12 of the 16 TrpB variants complemented growth to a similar degree as TRP5 when supplemented with 400 μM indole, demonstrating substantial improvement over their wt *Tm*TrpB and *Tm*Triple parents (Fig. 2.2a).



**b**

| OrthoRep-evolved TrpB variant | mutations |
|---|---|
| WT-100-1-A | G3D, Y28H, E29G, D34G, P54T, I69V, K96L, A118V, F122S, V245A, I271V, T292S, A351V |
| WT-003-1-A | A20V, A118V, F122S, V245A, I271V, T292S, A351V |
| Q90*-003-1-A | Y8H, *90Q, M97T, N167D, S243P, S302P, S335P |
| Q90*-003-1-B | Y8H, M18V, A20P, F57L, *90Q, M97T, N167D, S243P, S302P, S335P, A351V, Y383C |
| Tri-003-1-A | E15K, K33R, A58T, N167D, V245A, G253D, I271V, A351V, R385H |
| Tri-003-1-B | E15K, A58T, E63G, A118V, A119T, I184V, I271V, Y273C, I332V, S335P, A351V |
| Tri-003-2-A | E15K, A58T, A118V, E143Q, T179A, V250I, T259A, I271V, A351V |
| Tri-100-1-A | E22K, I88T, K95R, T117A, A119T, Y193H, A234V, I248V, I271V, K361R |
| Tri-100-2-A | Y8H, A119T, F122S, N167D, I174F, I195T, E206G, Y352C, I388T |
| Tri-100-2-B | E30G, K96L, P140L, N167D, I184F, L213P, G228S, I388T |
| Tri-100-3-A | K39E, K95_K96insK, T117A, K270E, I271N |
| Tri-100-4-A | R46H, F57Y, A118T, G123A, N167D, I241V, A330V, A349T, G360D, Y383H |
| Tri-100-4-B | G3D, E30G, I102T, P140L, N167D, L213P |

**Fig. 2.2. *In vivo* activity and diversity of individual *Tm*TrpB variants from OrthoRep-evolved populations. a** Evaluation of TRP5 complementation by evolved variants through a growth rate assay. Maximum growth rates over a 24-hour period for *Δtrp5* yeast strains transformed with a nuclear plasmid expressing the indicated *Tm*TrpB variant, grown in medium with or without 400 μM indole. Points and error bars represent mean ± s.d. for four biological replicates, respectively. Shaded area is the mean ± s.d. growth rate for the TRP5 positive control (*i.e.* plasmid expressing the endogenous yeast TRP5). Green box indicates the mean ± s.d. growth rate for all strains shown when Trp is supplemented. Growth rates for individual replicates in all three media conditions are shown in Fig. S2.2. Note that growth rates below ~0.15 per hour correspond to cultures that did not enter exponential phase; in these cases, the reported growth rate is not meaningful and instead can be interpreted as no quantifiable growth. **b** Parent populations from which OrthoRep-evolved variants shown in **a** are derived and all non-synonymous mutations present in each.

Unsurprisingly, this set of clonal *Tm*TrpBs contained more sequence diversity than the consensus sequences of the ten populations from which they were taken. Together, the variants tested comprised a total of 85 unique amino acid substitutions, with an average of 8.7 (± 2.1 s.d.) and a range of 5–13 non-synonymous mutations per variant (variant set 1, Tables S2 and S3). Since the 12 *Tm*TrpBs from this set exhibiting complementation were all similarly active in their primary activity yet mutationally diverse (Fig. 2.2b), we may conclude that our scaled evolution experiments generated substantial cryptic genetic variation. We note that four of 16 *Tm*TrpB variants exhibited similar or lower Trp productivity compared to their parent (Fig. S2.2). We suspect that the multicopy nature of p1 in the OrthoRep system allowed for deleterious mutations that appeared toward the end of the experiment to be maintained for a period of time without experiencing purifying selection if they arose in the same cell as functional variants, explaining the presence of these low activity *Tm*TrpBs. Indeed, this multicopy "buffering" may have worked to our advantage by promoting genetic drift under selection, facilitating both greater adaptation and greater diversity of evolutionary pathways across replicates (see **Discussion**). This may partly account for the high activity and high cryptic genetic variation present in the evolved *Tm*TrpBs.

### Evolved TmTrpBs exhibit high primary and promiscuous activity in vitro

We further characterized the evolved *Tm*TrpBs *in vitro* to approximate conditions of industrial application, make kinetic measurements, and test whether promiscuous activity could be detected. Nine *Tm*TrpB variants were sampled from those that supported robust indole-dependent growth in the Δ*trp5* strain, cloned into an *E. coli* expression vector with a C-terminal polyhistidine tag, and overexpressed. To mimic streamlined purification conditions compatible with biocatalytic application of *Tm*TrpBs, we generated heat treated *E. coli* lysates (1 hour incubation at 75 °C) and tested them for their ability to couple indole and serine to produce Trp at 30 °C. Three of the nine OrthoRep-evolved TrpBs, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A, demonstrated improved activity over the benchmark *Tm*Triple, as measured by total turnover number (TTN) (Fig. S2.3). Conveniently, each of these variants was evolved from a different starting point, meaning that wt *Tm*TrpB, *Tm*Triple, and *Tm*TripleQ90* were all viable starting points for reaching high activity *Tm*TrpBs. (See Table S2.3 for an explanation of variant naming conventions where each name designates the source of the variant. For example, WT-003-1-A designates variant A taken from the first replicate of a 3 mL evolution experiment starting from wt *Tm*TrpB.)

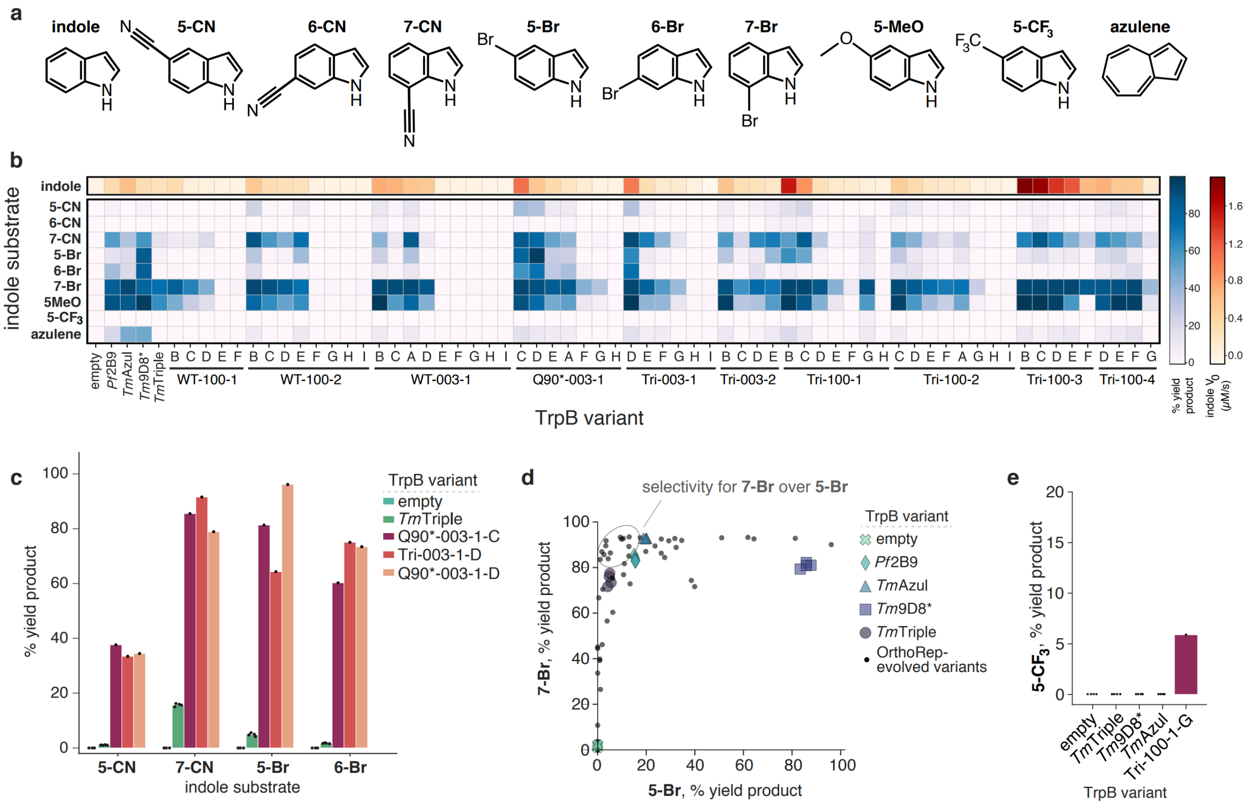


**Fig. 2.3. Promiscuous activities of a panel of evolved *Tm*TrpBs. a** Indole substrates used to test the substrate scope of a panel of *Tm*TrpB variants. **5-CN**, 5-cyanoindole; **6-CN**, 6-cyanoindole; **7-CN**, 7-cyanoindole; **5-Br**, 5-bromoindole; **6-Br**, 6-bromoindole; **7-Br**, 7-bromoindole; **5-MeO**, 5-methoxyindole; **5-CF$_3$**, 5-trifluoromethylindole. **b** Heatmap of TrpB activities reported as yield of the Trp analog produced from indicated substrates where 100% yield corresponds to full conversion of the indole analog to the Trp analog. Reactions were carried out using heat treated (1 hour at 75° C) cell lysate, yield was measured by HPLC-MS, and $V_0$ is the initial rate of Trp formation from indole at saturating serine concentrations. Panel *Tm*TrpB variants are ordered first by the parental cultures from which they were derived, then by activity with indole. Empty designates expression vector without any TrpB encoded. Reactions with OrthoRep-evolved variants other than WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A were performed in one replicate, empty performed in three biological replicates, and all other reactions performed in four biological replicates.. **c** Bar graph of selected indole analog activities from panel **b**. Points represent % HPLC yield for individual replicates, bars represent mean yield for multiple replicates or yield for a single replicate. **d** Activities of all variants shown in **b** for reactions with substrates **7-Br** and **5-Br** to show selectivity. Individual replicates are shown for empty vector and benchmark TrpBs and only mean values are shown for OrthoRep-evolved variants tested in

replicate (*i.e.* WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A) for clarity. **e** Heat treated lysate activity with **5-CF$_3$** for indicated TrpB variants.

Since the benchmark *Tm*Triple against which we compared the evolved *Tm*TrpBs was engineered through classical directed evolution involving screening *E. coli* lysates, whereas our *Tm*TrpB variants were evolved in yeast but expressed in *E. coli* for characterization, it is likely that the high-activity evolved *Tm*TrpBs would compare even more favorably if normalized by expression. We therefore purified WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A by immobilized metal affinity chromatography (IMAC) and reevaluated their activity for coupling indole with serine to generate Trp. By TTN, all three variants showed a 4- to 5-fold increase in activity over *Tm*Triple at 30 °C (Fig. S2.4). At 75 °C, however, WT-003-1-A had only ~2-fold higher activity than *Tm*Triple, while the other two variants were less active than *Tm*Triple. Since the thermostability of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A had not been reduced dramatically ($T_{50} > 83.7$ °C, Fig. S2.5), adaptation in these variants occurred at least partially by shifting the activity temperature profile. This is a practically valuable adaptation, since thermostable enzymes that operate at mesophilic temperatures allow for greater versatility in application without sacrificing durability and ease of purification through heat treatment.

Further testing of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A revealed that all three enzymes had at least a 22-fold higher $k_{cat}/K_M$ for indole than did *Tm*Triple at 30 °C (Table S2.4 and Fig. S2.6). Finally, testing for production of Trp analogs revealed that these variants' improved performance with indole transferred to alternate substrates (Fig. S2.4), validating their utility as versatile biocatalysts and also the hypothesis that continuous evolution of *Tm*TrpB variants can uncover promiscuous activities for which they were not selected.

## A diverse panel of evolved TmTrpB variants encompasses a variety of useful promiscuous activities with indole analogs

Given the exceptional performance of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A and their ability to transfer primary activity to new substrates as promiscuous activity, we decided to further sample the variant diversity generated across the multiple *Tm*TrpB evolution experiments. We cloned 60 randomly chosen *Tm*TrpBs from the ten continuous evolution populations into *E. coli* expression vectors for *in vitro* characterization. These 60 *Tm*TrpBs represent extensive diversity, with an average of 9.3 (± 2.8 s.d.) non-synonymous mutations per variant and a total of 194 unique amino acid changes across the set; in addition, each sequence encoded a unique protein (variant set 2, Tables S2 and S3). Since each variant had multiple non-synonymous mutations (up to 16) accumulated through >100 generations of adaptation and neutral drift, the depth of OrthoRep-based evolution was indeed leveraged in their evolution. We visualized these sequences, together with the consensus sequences of the populations from which they were derived, as nodes in a force directed graph related by shared mutations (Fig. S2.7). With only one exception, all individual sequences cluster near the consensus sequence for their population, meaning that interpopulation diversity exceeded intrapopulation diversity. Thus, the scale of OrthoRep-based evolution was also leveraged in these variants — if fewer independent evolution experiments had been run, the reduction in diversity would not be recoverable from sampling more clones.

Preparations of *Tm*TrpBs WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A, the 60 new variants, and four top-performing TrpB benchmark variants from past classical directed evolution campaigns (including *Tm*Triple) were all tested for product formation with indole by UV absorption and nine indole analogs by high performance liquid chromatography-mass spectrometry (HPLC-MS) to detect substrate promiscuity (Fig. 2.3a). The panel of 63 OrthoRep-evolved *Tm*TrpB variants exhibited an impressive range of activities (Fig. 2.3b). First, we observed that a number of variants had primary activities with indole that surpass the benchmark *Tm*Triple in lysate, with initial velocities of Trp formation up to 3-fold higher than WT-003-1-A (Fig. 2.3b and Fig. S2.8) whose

$k_{cat}/K_M$ for indole is 1.37 x $10^5$ $M^{-1}$ $s^{-1}$, already 28-fold higher than *Tm*Triple's at saturating serine concentrations (Table S2.4 and Fig. S2.6). Second, direct comparison of some of the best panel variants to *Tm*Triple revealed dramatic general activity improvements for multiple indole analogs (Fig. 2.3b). For example, across the three most versatile variants (Q90*-003-1-C, Tri-003-1-D, and Q90*-003-1-D) the maximum fold-improvement in product yields over *Tm*Triple were 37, 5, 19, and 50 using substrates **5-CN**, **7-CN**, **5-Br**, and **6-Br**, respectively (Fig. 2.3c). Finally, with the exception of **6-Br** and **azulene**, at least one variant from the OrthoRep-evolved panel converted the indole analog substrates as well as or better than benchmark TrpBs *Pf*2B9, *Tm*Azul, and *Tm*9D8*, which had been deliberately engineered toward new substrate scopes, though at higher temperatures (Figs. 3.3b and Fig. S2.9).[7,21,25,26]

The diverse properties represented in our 63 variants were not just limited to primary activity increases on indole and promiscuous activities for indole analogs. Multiple variants from the panel also exhibited substantial improvements in selectivity for differently substituted indoles, which could be useful when working with substrate mixtures that may be less expensive to use industrially. For example, we observed many *Tm*TrpBs with greater selectivity for **7-Br** over **5-Br** as compared to all four of the benchmark engineered TrpBs (Fig. 2.3d). Another variant in the panel, Tri-100-1-G, stood out for having appreciable activity with nearly all substrates tested, including **6-CN** and **5-CF$_3$**, which are poorly utilized by most other TrpBs, likely due to electron-withdrawing effects of their respective moieties. Notably, the ability to accept **5-CF$_3$** as a substrate was unique to Tri-100-1-G: all other variants, as well as the benchmark TrpBs, showed no detectable product formation with this substrate (Fig. 2.3e and Fig. S2.9). Repeating the reaction with purified enzyme in replicate confirmed the observed activity (Fig. S2.10). Tri-100-1-G may therefore be a promising starting point for new engineering efforts to access exotic Trp analogs. In short, despite having been selected for native activity with indole, OrthoRep-evolved *Tm*TrpBs have extensive and diverse activities on a range of non-native substrates, demonstrating the value of depth and scale in the evolution of enzyme variants.

### Mutations in evolved *Tm*TrpBs may modulate conformational dynamics and fine tune the active site

Of the ~200 unique mutations in the OrthoRep-evolved *Tm*TrpBs that we characterized, there were some mutations whose effects could be rationalized from comparison to previous work. Since the *Tm*TrpBs had to evolve stand-alone activity, it is unsurprising that many of the mutations we observed have been implicated in the loss of allosteric regulation by TrpA. For example, Buller *et al.* previously examined a series of engineered variants from *Pyrococcus furiosus* TrpB (*Pf*TrpB) and found that evolution for stand-alone activity was facilitated by a progressive shift in the rate-limiting step from the first to the second stage of the catalytic cycle as well as stabilization of the 'closed' conformation of the enzyme.[27] That work implicated eight residues in this mechanism, seven of which correspond to homologous sites where we observed mutations in the evolved *Tm*TrpB variants (*i.e.*, P14, M18, I69, K96, L274, T292, and T321). Another mutation, N167D, present in three of the ten consensus sequences for evolved populations (Fig. 2.1c), has also been implicated in stabilizing the closed state.[21] Additional mutations observed but not studied before (*e.g.*, S277F, S302P, and A321T) could also reasonably alter the allosteric network linking *Tm*TrpB activity to its natural *Tm*TrpA partner, based on existing structures and molecular dynamics simulations on the homologous *Pf*TrpA/*Pf*TrpB complex.[22,27] Taken together, these mutations are likely implicated in converting allosteric activation by *Tm*TrpA into constitutive activity to establish stand-alone function of *Tm*TrpBs.

During the evolution of stand-alone activity, not only must allosteric activation by *Tm*TrpA be recapitulated by mutations in *Tm*TrpB, the surface of *Tm*TrpB that normally interacts with *Tm*TrpA must adjust to a new local environment. Consistent with this adaptation, all consensus sequences for the ten successfully evolved populations from which our *Tm*TrpB variants were sampled contain a mutation to at least one of a set of five residues located on the canonical TrpA interaction interface (Fig. 2.1c and Fig. S2.11). These mutations might improve solubility by increasing hydrophilicity (*e.g.* G3D, Y8H, and A20T) or form new intramolecular interactions that compensate for lost interactions with *Tm*TrpA, among other possibilities.

We also detected strong convergent evolution in a region near the catalytic lysine, K83, which directly participates in *Tm*TrpB's catalytic cycle through covalent binding of PLP and multiple proton transfers (Fig. S2.12).[19] For example, A118 was mutated in the consensus sequence of four of the ten fully adapted populations, while adjacent residues T117 or A119 were mutated in an additional three (Fig. 2.1c). Furthermore, the three populations in which these residues were not mutated contained other consensus mutations that are either part of the α-helix to which K83 belongs, or, like residues 117–119, within ~8 Å of this helix (Fig. 2.1c and Fig. S2.12). We hypothesize that the α-helix harboring K83 is a focal point of evolution, whereby mutations in its vicinity may finely adjust the positioning of K83 and the PLP cofactor to improve catalysis, perhaps as compensation for structural changes induced by thermoadaptation. Some OrthoRep-evolved variants also contained mutations to first- and second-shell active site residues (Fig. S2.13), which may directly modulate the activity of *Tm*TrpBs, although these mutations were rare. Taken together, we hypothesize that these mutations near the active site residues of TrpB were adaptive or compensatory.

The ~20 mutations considered above are rationalized with respect to their impact on *Tm*TrpB's primary catalytic activity. While substrate promiscuity changes may be influenced by these explainable mutations, previous literature suggests that substrate specificity is globally encoded by amino acids distributed across an entire enzyme.[28] Indeed, the majority of the ~200 mutations found in our panel of *Tm*TrpBs were far away from *Tm*TrpB's active site and not rationalizable based on the known structural and kinetic properties of TrpBs. We suspect that the cryptic genetic variation this majority of mutations encompasses contributes to the diversity in substrate scope across our variants.

## 2.4 Discussion

In this work, we showed how the depth and scale of evolutionary search available in OrthoRep-driven protein evolution experiments could be applied to broaden the secondary promiscuous activities of *Tm*TrpB while only selecting on its primary activity. The significance of this finding can be divided into two categories, one concerning the practical utility of the new *Tm*TrpB variants we obtained and the second concerning how this evolution strategy may apply to future enzyme evolution campaigns and protein engineering in general.

Practically, the new *Tm*TrpBs should find immediate use in the synthesis of Trp analogs. Trp analogs are valuable chiral precursors to pharmaceuticals as well as versatile molecular probes, but their chemical synthesis is challenged by stereoselectivity requirements and functional group incompatibility. This has spurred enzyme engineers to evolve TrpB variants capable of producing Trp analogs,[20,21,25,26] but the capabilities of available TrpBs are still limited. Compared to existing engineered TrpBs, our new panel of variants has substantially higher activity for the synthesis of Trp and Trp analogs at moderate temperatures from almost all indole analogs tested and also accepts indole analogs, such as **5-CF$_3$** (Fig. 2.3a), for which benchmark TrpBs used in this study showed no detectable activity (Fig. 2.3e). (In fact, only one TrpB variant has shown detectable activity for this substrate in previous classical directed evolution campaigns.[21]) In addition, at least one member of the panel accepted each of the nine indole analogs we used to profile promiscuity, suggesting that additional indole analogs and non-indole nucleophiles not assayed here will also be accepted as substrates.[29,30] Finally, the evolved *Tm*TrpBs are both thermostable and adapted for enzymatic activity at 30 °C. This maximizes their industrial utility, as thermostability predicts a protein's durability and can be exploited for simple heat-based purification processes, while mesophilic activity is compatible with heat-labile substrates, industrial processes where heating costs can compound, or *in vivo* applications in model mesophilic hosts (*e.g. S. cerevisiae* or *E. coli*).

Of more general significance may be the process through which the *Tm*TrpBs in this study were generated. Previous directed evolution campaigns aiming to expand the substrate scope of TrpB screened directly for activity on indole analogs to guide the evolution process,[21,26] whereas this study only selected for *Tm*TrpB's primary activity on indole. Yet this study still yielded *Tm*TrpBs whose secondary activities on indole analogs were both appreciable and diverse. Why?

A partial explanation may come from the high primary activities of OrthoRep-evolved *Tm*TrpBs, as validated by kinetic measurements showing that variants tested have $k_{cat}/K_M$ values for indole well in the $10^5$ $M^{-1}$ $s^{-1}$ range. Since OrthoRep drove the evolution of *Tm*TrpB in a continuous format for >100 generations, each resulting *Tm*TrpB is the outcome of many rounds of evolutionary improvement and change (evolutionary depth). This contrasts with previous directed evolution campaigns using only a small number of manual rounds of diversification and screening. Continuous OrthoRep evolution, on the other hand, allowed *Tm*TrpBs to become quite catalytically efficient with minimal researcher effort. We suggest that the high primary catalytic efficiencies also elevated secondary activities of *Tm*TrpB, resulting in the efficient use of indole analogs. However, this explanation is not complete, as evolved *Tm*TrpBs with similar primary activity on indole had differences in secondary activities (Fig. 2.3). In other words, high primary activities did not uniformly raise some intrinsic set of secondary activities in *Tm*TrpB, but rather influenced if not augmented the secondary activities of *Tm*TrpB in different ways. We attribute this to the fact that we ran our evolution experiments in multiple independent replicates (evolutionary scale). Each replicate could therefore evolve the same primary activity through different mutational paths, the idiosyncrasies of which manifest as distinct secondary activities. A third explanation for the promiscuous profile diversity of these *Tm*TrpB variants is that each replicate evolution experiment had, embedded within it, mechanisms to generate cryptic genetic variation without strong selection on primary activity. Many of the clones we sampled from each *Tm*TrpB evolution experiment had novel promiscuity profiles but mediocre primary activity with indole (Fig. 2.3). We believe this is because OrthoRep drove *Tm*TrpB evolution in the context of a multicopy plasmid such that non-neutral genetic drift from high activity sequences could occur within each cell at any given point. Therefore, *Tm*TrpB sequences with fitness-lowering mutations could persist for short periods of time, potentially allowing for the crossing of fitness valleys during evolution experiments and, at the end of each evolution experiment, a broadening of the genetic diversity of clones even without explicitly imposed periods of relaxed selection. Since enzyme orthologs are capable of specializing towards different sets of secondary activities when pressured to do so,[2,3] non-neutral genetic drift from different consensus sequences across independent population should also access different secondary activities, further explaining the diversity of promiscuous activity profiles across clones selected from replicate evolution experiments. The combination of these mechanisms likely explains the variety of properties encompassed by the panel of *Tm*TrpBs.

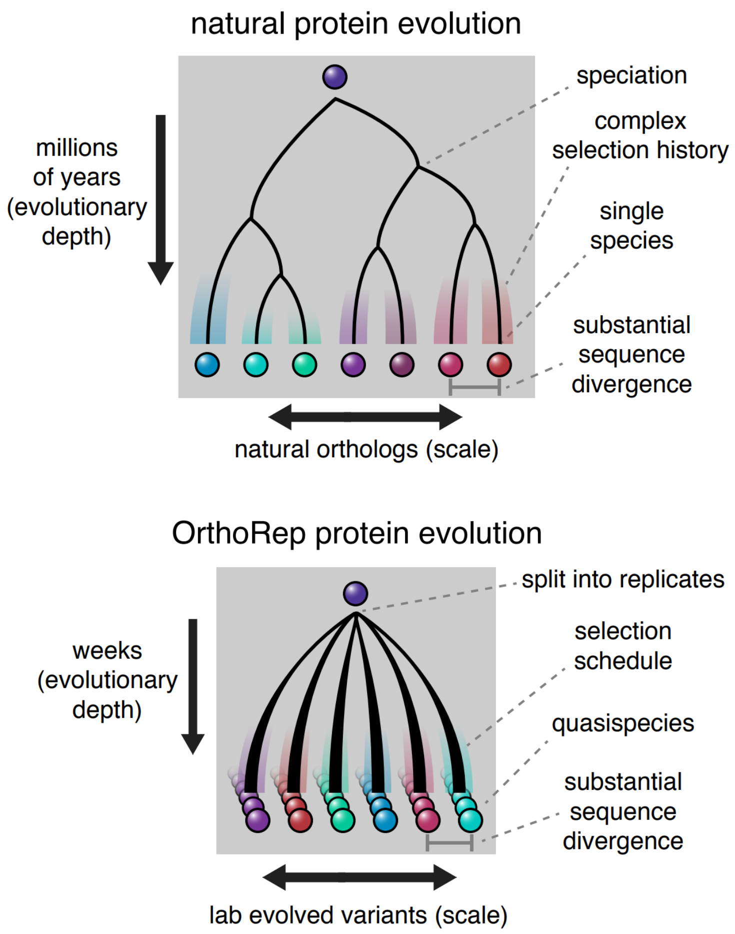

natural protein evolution
millions of years (evolutionary depth)
speciation
complex selection history
single species
substantial sequence divergence
natural orthologs (scale)
OrthoRep protein evolution
weeks (evolutionary depth)
split into replicates
selection schedule
quasispecies
substantial sequence divergence
lab evolved variants (scale)

**Fig. 2.4. Conceptual similarities between natural enzyme ortholog evolution and OrthoRep evolution.** Splitting OrthoRep cultures into many replicates can be seen as a form of speciation occurring by spatial separation. Complex selection schedules may emulate varied selection histories of natural orthologs, generating substantial sequence divergence across replicates. Evolved OrthoRep cultures contain diverse populations of sequences akin to quasispecies owing to high mutation rates.[41]

Our approach to *Tm*TrpB evolution was inspired by the idea of gene orthologs in nature. Orthologs typically maintain their primary function while diversifying promiscuous activities through long evolutionary histories in different species.[2,31] We approximated this by evolving *Tm*TrpB through continuous rounds of evolution, mimicking long histories, and in multiple replicates, mimicking the spatial separation and independence of species. Such depth and scale of evolutionary search is likely responsible for the substrate scope diversity of the *Tm*TrpBs we report even as they were selected only on their primary activity. We recognize that the evolved *Tm*TrpBs represent lower diversity than natural orthologs. For example, the median amino acid sequence divergence between orthologous human and mouse proteins is 11%,[32] while the median divergence between pairs of variants from our experiment is 4.3% with a maximum of 8% (Fig. S2.14). Still, this level of divergence between functional variants is substantial for a laboratory protein evolution experiment and suggests that it is realistic to model future work on the processes of natural ortholog evolution (Fig. 2.4). For example, it should be straightforward to scale our experiments further, to hundreds or thousands of independent populations each evolving over longer periods of time. This would better simulate the vastness of natural evolution. It should also be possible to deliberately vary selection schedules by adding competitive *Tm*TrpB inhibitors (such as the very indole analogs for which they have promiscuous activity), changing temperatures, or cycling through periods of weak and strong selection at different rates. Such evolutionary courses would approximate complexity in natural evolutionary histories. These modifications to OrthoRep-driven *Tm*TrpB evolution should yield greater cryptic genetic diversity, which may result in further broadening of promiscuous functions. The generation of cryptic genetic diversity at depth and scale should also be useful in efforts to predict protein folding and the functional effects of mutations via co-evolutionary analysis.[33,34] Indeed, catalogs of natural orthologs have proven highly effective in fueling such computational efforts, so our ability to mimic natural ortholog generation on laboratory timescales may be applicable to protein biology at large. Within the scope of enzyme engineering, we envision that the process of continuous replicate evolution, selecting only on primary activities of enzymes, will become a general strategy for expanding promiscuous activity ranges of enzymes as we and others extend it to new targets.

## 2.5 Materials and Methods

### DNA plasmid construction

All plasmids that were not generated in a previous study were constructed via Gibson assembly[35] from parts derived from the Yeast Toolkit,[24] from previously described OrthoRep integration cassette plasmids,[16] from *E. coli* expression vectors for previously described TrpB variants,[7,26] from synthesized oligonucleotides, from yeast genomic DNA, or from the standard *E. coli* expression vector, pET-22b(+). All DNA cloning steps and *E. coli* protein expression steps were performed in *E. coli* strains TOP10 and BL21(DE3), respectively. All oligonucleotides used for PCR were purchased from IDT, and all enzymes and reagents used for cloning were purchased from NEB.

Parts used to generate yeast nuclear expression plasmids for testing the selection and p1 integration plasmids were PCR amplified from DNA sources listed above, Gibson assembled, transformed into *E. coli*, and plated onto selective LB agar plates. Individual clones were picked, grown to saturation in selective LB liquid media, miniprepped, and sequence confirmed. Following evolution of TrpB, individual variants were assembled into new yeast or *E. coli* expression vectors through PCR amplification of purified DNA from evolved yeast cultures, bulk cloning into the appropriate expression vector, picking individual colonies, and confirming absence of any frameshift mutations by Sanger sequencing.

### Yeast strains and media

Yeast were incubated at 30 °C, with shaking at 200 rpm for liquid cultures, and were typically grown in synthetic complete (SC) growth medium (20 g/L dextrose, 6.7 g/L yeast nitrogen base w/o amino acids (US Biological), 2 g/L SC dropout (US Biological) minus nutrients required for appropriate auxotrophy selection(s)), or were grown in YPD growth medium (10 g/L bacto yeast extract, 20 g/L bacto peptone, 20 g/L dextrose) with or without antibiotics, if no auxotrophic markers were being selected for. Media agar plates were made by combining 2X concentrate of molten agar and 2X concentrate of desired media formulation. Prior to all experiments, cells were grown to saturation in media selecting for maintenance of any plasmids present.

### Yeast transformation

All yeast transformations were performed as described in Gietz and Shiestl.[36] Briefly, a 4 mL culture of yeast was grown to mid-log phase in rich YPD medium (2% (w/v) Bacto yeast extract, 4% (w/v) Bacto peptone, 4% (w/v) glucose) at 30 °C, harvested by centrifugation, washed with sterile water, and pelleted again by centrifugation. These pellets were then resuspended in a mixture containing PEG3350 (30% (w/v) final concentration), lithium acetate (90 mM final concentration), boiled salmon sperm carrier DNA (0.25 mg/mL final concentration), and ~1 μg of the DNA to be transformed, all in a total volume of 410 μL. Transformations were often done in smaller scales where the described volumes were split into 8 transformations. Cells resuspended in the PEG3350/lithium acetate/DNA mixture were then incubated at 42 °C for 30 min., pelleted by centrifugation, resuspended in 1 mL YPD medium, incubated for 1 hour at 30 °C with shaking (200 rpm), pelleted, resuspended in 1 mL sterile water, pelleted, and resuspended in an appropriate volume of sterile water or 0.9% (w/v) NaCl for plating. Transformed cells were streaked onto selective media agar plates, and resulting single colonies were picked for all further uses. Transformations for integration onto p1 were performed as described previously:[15] 2–4 μg of plasmid DNA with ScaI restriction sites adjacent to integration flanks was cut with ScaI-HF (NEB) and transformed into yeast harboring the wt p1 and p2 plasmids. Proper integration was validated by miniprepping the resulting clonal strain, visualizing the recombinant p1 band of the desired size by gel electrophoresis of the miniprepped DNA, and PCR and Sanger sequencing of the gene of interest integrated onto p1.

To generate enough DNA for visualization of the recombinant p1 plasmid, high yield yeast minipreps were performed as previously described.[15] In brief, 1.5 mL of culture was pelleted, supernatant was discarded, and the pellet was resuspended in 1 mL 0.9% NaCl, pelleted again, and resuspended in 250 µL Zymolyase solution (0.9 M D-Sorbitol (Sigma Aldrich), 0.1 M Ethylenediaminetetraacetic acid (EDTA, Sigma Aldrich), 10 U/mL Zymolyase (US Biological)). The suspension was incubated at 37 °C for 1 hour, then centrifuged at 3,000*g* for 5 min. The supernatant was discarded, and the pellet was resuspended in 280.5 µL proteinase K solution (250 µL TE (50 mM Tris-HCl (pH 7.5), 20 mM EDTA), 25 µL 10% sodium dodecyl sulfate (SDS, Sigma Aldrich), 5.5 µL proteinase K stock solution (10 mg/mL proteinase K (Fisher) in water). Samples were then incubated at 65 °C for 30 min, combined with 75 µL 5 M potassium acetate (Fisher), and incubated on ice for 30 min. Samples were centrifuged at 12,000*g* for 10 min, the resulting supernatant was combined and mixed with 700 µL ethanol (Gold Shield), and samples were then centrifuged at 3,500*g* for 15 min. Supernatant was discarded and the resulting pellet was dried, resuspended in 150 µL TE, and centrifuged at 12,000*g* for 10 min. Supernatant was then combined with 8 µL 1 mg/mL ribonuclease A (Thermo Scientific) and incubated at 37 °C for 30 min, combined with 150 µL isopropanol (Fisher), and centrifuged at 12,000*g* for 10 min to pellet purified DNA. Pellet is dried, then resuspended in 30 µL water.

Following confirmation of the presence of the desired recombinant p1, strains were then transformed with either of two plasmids for nuclear expression of an OrthoRep terminal protein DNA polymerase 1 (TP-DNAP1) variant: wt TP-DNAP1 (pAR-Ec318) for evaluating *trp5* complementation of TrpB variants without mutagenesis, or error-prone TP-DNAP1 (pAR-Ec633) for generating strains ready for TrpB evolution. These strains were passaged for ~40 generations to stabilize copy number of the recombinant p1 species, prior to any use in experiments.

Genomic deletion of the entire *TRP5* ORF was accomplished through co-transformation of a CRISPR/Cas9 plasmid targeting *TRP5* with the spacer sequence TTTGAGCCTGATCCCACTAG and a linear DNA fragment comprised of two concatenated 50 bp homology flanks to the *TRP5* ORF, generated from primers X and X.[37] Transformations were then plated on selective media agar, colonies were re-streaked onto nonselective media agar, and resulting colonies were grown to saturation in liquid media. The region of interest was PCR amplified and Sanger sequenced to confirm presence of desired modification.

## Plating assays

Yeast strains expressing a TrpB variant either from a nuclear plasmid, or from p1 with wt OrthoRep polymerase (TP-DNAP1) expressed from a nuclear plasmid, were grown to saturation in SC –L or SC –LH, spun down, washed once with 0.9% NaCl, then spun down again, and the resulting pellet was resuspended in 0.9% NaCl. Washed cells were then diluted 1:100 (or 1:10,000 where indicated) in 0.9% NaCl, and 10 µL of each diluted cell suspension was plated onto media agar plates in pre-marked positions. After 3 days of growth, cell spots were imaged (Bio-Rad ChemiDoc™). Resulting images were adjusted uniformly ('High' set to 40,000) to improve visibility of growth (Bio-Rad Image Lab™ Software). Figures utilizing these images (Figs. S1 and S2) were made by manually combining images of different plates, but all images of the same media condition within each figure panel were derived from the same image of a single plate.

## *Tm*TrpB evolution

Yeast strains with a nuclear plasmid expressing error-prone TP-DNAP1 and with wt *Tm*TrpB, *Tm*Triple, or *Tm*TripleQ90* encoded on p1 (GR-Y053, GR-Y055, and GR-Y057) were grown to saturation in SC –LH, prior to passaging for evolution. All cultures passaged for evolution of *Tm*TrpB regardless of success are described in Table S2.1. To provide enough indole substrate for sufficient Trp production, but not enough to induce toxicity, all growth media used for evolution of TrpB activity was supplemented with 100 µM indole, as informed by results shown in Fig. S2.1. All passages for evolution were carried out as 1:100 dilutions. To induce a growth defect but still allow for some growth, the first passage for each evolution culture was carried out in SC –LH media with 37 µM Trp (7.6 mg/L). After two or three days of shaking incubation, if $OD_{600}$ > 1.0 (Bio-Rad SmartSpec™ 3000) for 100 mL cultures, or if most wells in a 24 well block of 3 mL cultures were saturated to a similar degree by eye, cultures were passaged into fresh growth medium with a slightly reduced Trp concentration. If the level of growth was beneath this threshold, the culture was passaged into growth medium with the same Trp concentration. This process was continued until cultures were capable of growth in a Trp concentration of 3.7 µM (or, in the sole case of WT-100-1, 4.7 µM), at which point a passage into media lacking Trp was attempted, which typically resulted in successful growth. Resulting cultures were then passaged six additional times into growth medium lacking Trp.

## Growth rate assays

Yeast strains containing nuclear plasmids encoding one of several OrthoRep-evolved TrpB variants, wt *Tm*TrpB, *Tm*Triple, or none of these (denoted ‘empty’) were grown to saturation in SC –L, washed as described above, then inoculated 1:100 into multiple media conditions in 96-well clear bottom plates, with four biological replicates per media/strain combination. Plates were then sealed with a porous membrane and allowed to incubate with shaking at 30 °C for 24 hours, with $OD_{600}$ measurements taken automatically every 30 minutes (Tecan Infinite M200 Pro), according to a previously described protocol.[38] Multiple 24 hour periods were required for each experiment, but empty controls were included in each individual 96-well plate to ensure validity of growth in other cultures. Raw $OD_{600}$ measurements were fed into a custom MATLAB script,[18] which carries out a logarithmic transformation to linearize the exponential growth phase, identifies this growth phase, and uses this to calculate the doubling time (*T*). Doubling time was then converted to growth rate (gr) using equation (1):

$$\mathrm{gr} = \frac{\ln(2)}{T}$$

## Enzyme characterization — general experimental methods

Chemicals and reagents were purchased from commercial sources and used without further purification. All cultures were grown in Terrific Broth supplemented with 100 µg/mL carbenicillin ($TB_{carb}$). Cultures were shaken in a New Brunswick Innova 4000 (shaking diameter 19 mm), with the exception of the 96-well deep-well plates (USA Scientific), which were shaken in a Multitron INFORS HT (shaking diameter 50 mm). Lysis buffer was composed of 50 mM potassium phosphate, pH 8.0 (KPi buffer), supplemented with 100 or 200 µM pyridoxal 5'-phosphate (PLP). Heat lysis was performed in a 75 °C water bath (Fisher) for >1 hour. Protein concentrations were determined using a Pierce™ BCA Protein Assay Kit (Thermo Scientific). Reactions were performed in KPi buffer. Liquid chromatography/mass spectrometry (LCMS) was performed on an Agilent 1290 UPLC-LCMS equipped with a C-18 silica column (1.8 µm, 2.1 × 50 mm) using $CH_3CN/H_2O$ (0.1% acetic acid by volume): 5% to 95% $CH_3CN$ over 2 min; 1 mL/min. Liquid chromatography/mass spectrometry (LCMS) was also performed on an Agilent 1260 HPLC-

MS equipped with Agilent InfinityLab Poroshell 120 EC-C18 column (2.7 μm, 4.6×50 mm): hold 5% $CH_3CN$ for 0.5 min, 5-95% $CH_3CN$ over 2 min; 1 mL/min.

TrpB variants selected for further characterization were cloned into a pET-22b(+) vector with a C-terminal 6X His-tag and transformed into *E. coli* BL21(DE3) cells (Lucigen).

### Expression and characterization of variants from set 1 — large scale expression and lysis

A single colony containing the appropriate TrpB gene was used to inoculate 5 mL $TB_{carb}$ and incubated overnight at 37 °C and 230 rpm. For expression, 0.5 mL of overnight culture were used to inoculate 50 mL $TB_{carb}$ in a 250 mL flask and incubated at 37 °C and 250 rpm for 3 hours to reach $OD_{600}$ 0.6–0.8. Cultures were chilled on ice for 20 min and expression was induced with a final concentration of 1 mM isopropyl β-D-thiogalactopyranoside (IPTG). Expression proceeded at 25 °C and 250 rpm for approximately 20 hours. Cells were harvested by centrifugation at 5,000*g* for 5 min at 4 °C, and then the supernatant was decanted. The pellet was stored at −20 °C until further use or used immediately for whole cell transformations.

Pellets were lysed in 5 mL of lysis KPi buffer with 200 μM PLP, supplemented with 1 mg/mL lysozyme (HEWL, Sigma Aldrich), 0.02 mg/mL bovine pancreas DNase I, and 0.1X BugBuster (Novagen) and incubated at 37 °C for 30 minutes. Lysate was clarified by centrifugation at 5,000*g* for 10 min, divided into 1 mL aliquots, and stored at −20 °C until further use.

### Expression and characterization of variants from set 1 — lysate and whole cell small-scale reactions

Protein concentration in lysate was quantified by BCA. Lysate reactions were performed in 2 mL glass HPLC vials (Agilent) charged with indole (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of lysate (final enzyme conc. 4 μM), and serine (final conc. 20 mM) to achieve a final volume of 200 μL. Whole cell reactions were performed in 2 mL glass HPLC vials (Agilent) charged with indole (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of cells diluted in KPi buffer (final $OD_{600}$=6), and serine (final conc. 20 mM) to achieve a final volume of 200 μL. Reactions were incubated at 30 °C for 24 hours, diluted with 800 μL 1:1 $CH_3CN$/1 M aq. HCl, and analyzed via UHPLC-MS.

## Expression and characterization of variants from set 1 — thermostability determination

Enzyme $T_{50}$ measurements (the temperature at which 50% of the enzyme is irreversibly inactivated after a 1 hour incubation) were used to report on the thermostability of the enzyme. In a total volume of 100 μL, samples were prepared in KPi buffer with 1 μM enzyme in PCR tubes and either set aside (25 °C) or heated in a thermal cycler on a gradient from 79–99 °C (OrthoRep-generated variants), or 59–99 °C (*Tm*Triple), for 1 hour, with each temperature performed in duplicate. Precipitated protein was pelleted via centrifugation and 75 μL of each sample was carefully removed and added to the wells of a 96-well UV-transparent assay plate containing 0.5 mM indole and 0.5 mM serine. Relative product formation was observed by measuring the change in absorbance at 290 nm to determine the temperature at which the sample had 50% residual activity compared to the 25 °C samples (modeled as a logistic function).

## Expression and characterization of variants from set 1 — enzyme kinetics

Enzymatic parameters, $k_{cat}$ and $K_{M,}$ for the conversion of indole to Trp were estimated via Bayesian inference assuming Michaelis-Menten behavior under saturating serine (40 mM) in KPi buffer. Briefly, initial velocities (*v*) were determined by monitoring Trp formation in a Shimadzu UV-1800 spectrophotometer at 30 °C for 1 min over a range of indole concentrations at 290 nm using the reported indole-Trp difference in absorbance coefficient ($\Delta\varepsilon_{290}$ = 1.89 mM$^{-1}$ cm$^{-1}$).[39] These velocities were modeled using equation (2):

$$v = \frac{V_{\mathrm{max}}[\mathrm{indole}]}{K_{\mathrm{M}} + [\mathrm{indole}]}$$

and estimates for *v* and $V_{\mathrm{max}}$ were converted to *k* and $k_{\mathrm{cat}}$ by normalizing for enzyme concentration. Parameter estimates are obtained as Hamiltonian Markov chain Monte Carlo (MCMC) posterior samples and reported as the median with their 95% credible regions (CR). The MCMC software used for sampling was Stan (pystan version 2.19.0.0). The sampling was performed with four separate chains, each starting with 2000 warm-up (disregarded) steps followed by 12000 posterior sampling steps. The priors chosen for $k_{\mathrm{cat}}$ and $K_{\mathrm{M}}$ were lognormal distributions with means log(150) and log(500), standard deviations 2.5 and 1.5, with units of sec$^{-1}$ and μM, respectively. This provided non-negative probability density that covered low-to-moderate values of the parameters for many known enzymes, but still had significant density out to very high values for each parameter. (In all cases, the data was shown to significantly inform the prior.) The code used to generate these estimates (along with example data) can be found at http://github.com/palmhjell/bayesian_kinetics.

## Expression and characterization of variants from set 2 — small-scale expression and lysis

Variants were arrayed into a 96-well deep-well plate along with *Tm*Triple, *Tm*9D8*, *Tm*AzuI, and *Pf*2B9. Individual colonies were grown in 600 μL TB$_{\mathrm{carb}}$ in 96-well polypropylene plates overnight at 37 °C, 250 rpm, 80% humidity. The following day, 20 μL of overnight culture was used to inoculate 630 µL $TB_{carb}$ in deep-well 96-well plates and grown at 37 °C, 250 rpm. After 4 hours, cultures were chilled on ice for 20–30 min and protein expression was induced with 50 µL IPTG (final conc. 1 mM) diluted in $TB_{carb}$. Cultures were shaken at 20 °C, 250 rpm for 20–24 hours, after which they were subjected to centrifugation at 5,000*g* for 10 min. The cell pellets were frozen at −20 °C until further use or used immediately.

## Expression and characterization of variants from set 2 — indole rate measurements

Pellets were lysed in either 600 µL of KPi buffer with 100 µM PLP and heat treated at 75 °C for 1 hour, or in 600 µL of this buffer supplemented with 1 mg/mL lysozyme, 0.02 mg/mL bovine pancreas DNase I, and 0.1X BugBuster and incubated at 37 °C for 1 hour. Lysate from both conditions was clarified by centrifugation at 4,500*g* for 10 min and stored at 4 °C until further use.

Reaction master mix composed of 625 µM indole and 25 mM serine in KPi buffer was prepared and, before reactions, plates and master mix were incubated in 30 °C water bath for 30 min. The microplate reader (Tecan Spark) was also pre-heated to 30 °C.

To UV-transparent 96-well assay plates (Caplugs, catalog # 290-8120-0AF), 160 µL pre-heated reaction master mix was added by 12-channel pipet followed by 40 µL of lysate from the pre-heated plate using a Microlab NIMBUS96 liquid handler (Hamilton). Plates were immediately transferred into the plate reader, shaken for 10 sec to mix and the absorbance of each well at 290 nm was recorded as rapidly as possible (~20 sec between measurements) for 120 cycles. The rate of product formation was determined by finding the rate of absorbance change over time and converting to units of concentration using $\Delta\varepsilon_{290}$ = 1.89 $mM^{-1}$ $cm^{-1}$ (see above) and a determined path length of 0.56 cm. We observed no systematic difference in activity between the two lysate preparations (Fig. S2.15), suggesting that most enzyme variants retained sufficient thermostability for purification via heat treatment, and this method was used in subsequent experiments.

## Expression and characterization of variants from set 2 — substrate scope screen

Pellets were lysed in 300 µL KPi buffer with 200 µM PLP and clarified by centrifugation at 4,000*g* for 10 min. To a 96-well deep-well plate charged with 10 µL nucleophile dissolved in DMSO (See Table 1 for final reaction concentrations), 40 µL of the heat treated lysate was transferred using a Microlab NIMBUS96 liquid handler (Hamilton), followed by addition of 150 µL serine (final conc. 20 mM) with a 12-channel pipet. Reactions were sealed with 96-well ArctiSeal™ Silicone/PTFE Coating (Arctic White) and incubated in 30 °C water bath for ~24 hours. Reactions were diluted with 600 µL 2:1 $CH_3CN$/1 M aq. HCl, subjected to centrifugation at 5,000*g*, and 400 µL was transferred to 2 mL glass HPLC vials (Agilent). Samples were analyzed by HPLC-MS. Azulene samples were further diluted 20X to avoid oversaturation of the UV-detector and analyzed via UHPLC-MS.

All samples except those containing azulene were analyzed at 277 nm, representing the isosbestic point between indole and Trp and allowing estimation of yield by comparing the substrate and product peak areas for indole analogs.[21] Azulene yield was estimated as described previously.[25] Nucleophile retention times were determined though injection of authentic standards and product retention times were identified by extracting their expected mass from the mass spectrum.

## Large-scale expression and purification of Tri-100-3-F and Tri-100-1-G

A single colony containing the appropriate TrpB gene was used to inoculate 5 mL $\mathrm{TB_{carb}}$ and incubated overnight at 37 °C and 230 rpm. For expression, 2.5 mL of overnight culture were used to inoculate 250 mL $\mathrm{TB_{carb}}$ in a 1-L flask and incubated at 37 °C and 250 rpm for 3 hours to reach $\mathrm{OD_{600}}$ 0.6–0.8. Cultures were chilled on ice for 20 min and expression was induced with a final concentration of 1 mM IPTG. Expression proceeded at 25 °C and 250 rpm for approximately 20 hours. Cells were harvested by centrifugation at 5,000*g* for 5 min at 4 °C, and then the supernatant was decanted. The pellet was stored at −20 °C until further use.

Pellets were lysed in 25 mL KPi buffer with 200 μM PLP for >1 hour at 75 °C. Lysate was clarified by spinning 14,000*g* for 20 min at 4 °C (New Brunswick Avanti J-30I). Protein was purified over hand-packed HisPur™ Ni-NTA Resin (Thermo Scientific, catalog # 88221), dialyzed into KPi buffer and quantified by BCA.

### Tri-100-3-F PLP-binding assay

Variant Tri-100-3-F did not the exhibit characteristic yellow color of PLP-bound TrpB variants after purification, however BCA indicated comparable protein concentrations to the Tri-100-1-G variant. We have previously observed that some TrpB variants lose binding affinity for PLP resulting in non-functional apoenzyme. We evaluated Trp formation of Tri-100-3-F supplemented with 0, 0.1, 0.25, 0.5, 1, 2, 5, and 100 μM PLP via UV-Vis spectrophotometry. Serine (final conc. 25 mM) + PLP master mixes of the eight concentrations were prepared and dispensed into 96-well UV-transparent plate. Enzyme (final conc. 1 μM) with or without indole master mixes were prepared and 100 μL dispensed into 96-well plate. The plate was immediately transferred into plate reader, shaken for 10 sec to mix and product formation was measured ~20 sec for 120 cycles at 290 nm.

Only the 100 µM condition restored activity, supporting our hypothesis that the purified enzyme was apoprotein and binds PLP poorly, requiring supplementation of PLP to re-form a functional holoenzyme. Thus, we chose to supplement PLP in the subsequent purified protein reactions.

### Tri-100-3-F and Tri-100-1-G small-scale analytical reactions

Reactions were performed in 2 mL glass HPLC vials (Agilent) charged with nucleophile (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of purified protein (final enzyme conc. either 2 µM or 40 µM), PLP (final conc. 100 µM) and serine (final conc. 20 mM) to achieve a final volume of 200 µL. Reactions were incubated at 30 °C for ~24 hours. Reactions were diluted with 600 µL 2:1 $CH_3CN$/1 M aq. HCl, subjected to centrifugation at 5,000*g*, and 400 µL transferred to 2 mL glass HPLC vials (Agilent). Samples were analyzed by HPLC-MS. Azulene samples were further diluted 20X to avoid oversaturation of the UV-detector and analyzed via UHPLC-MS.

## 2.6 References

1. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science.* **278**, 631–637 (1997).
2. Khanal, A., McLoughlin, S. Y., Kershner, J. P. & Copley, S. D. Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. *Mol. Biol. Evol.* **32**, 100–108 (2015).
3. Baier, F. *et al.* Cryptic genetic variation shapes the adaptive evolutionary potential of

enzymes. *Elife* **8**, 1–20 (2019).

4. Aharoni, A. *et al.* The ‘evolvability’ of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).

5. Tawfik, O. K. and D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).

6. Leveson-Gower, R. B., Mayer, C. & Roelfes, G. The importance of catalytic promiscuity for enzyme design and evolution. *Nat. Rev. Chem.* **3**, 687–705 (2019).

7. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A Panel of TrpB Biocatalysts Derived from Tryptophan Synthase through the Transfer of Mutations that Mimic Allosteric Activation. *Angew. Chemie - Int. Ed.* **55**, 11577–11581 (2016).

8. Devine, P. N. *et al.* Extending the application of biocatalysis to meet the challenges of drug development. *Nat. Rev. Chem.* **2**, 409–421 (2018).

9. Truppo, M. D. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. *ACS Med. Chem. Lett.* **8**, 476–480 (2017).

10. Almhjell, P. J., Boville, C. E. & Arnold, F. H. Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997 (2018).

11. Zheng, J., Payne, J. L. & Wagner, A. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science.* **365**, 347–353 (2019).

12. Gupta, R. D. & Tawfik, D. S. Directed enzyme evolution via small and effective neutral drift libraries. **5**, 939–942 (2008).

13. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2**, 7–10 (2007).

14. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).

15. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

16. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell* **175**, 1946-1957.e13 (2018).

17. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

18. Zhong, Z. *et al.* Automated continuous evolution of proteins in vivo. *ACS Synth. Biol.* (2020) doi:10.1021/acssynbio.0c00135.

19. Dunn, M. F. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Arch. Biochem. Biophys.* **519**, 154–166 (2012).

20. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci.* **112**, 14599–14604 (2015).

21. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking Reactivity of TrpB: A General Biocatalytic Platform for Synthesis of Tryptophan Analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

22. Maria-Solano, M. A., Iglesias-Fernández, J. & Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **141**, 13049–13056 (2019).

23. Zhong, Z., Ravikumar, A. & Liu, C. C. Tunable Expression Systems for Orthogonal DNA Replication. *ACS Synth. Biol.* **7**, 2930–2934 (2018).

24. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).

25. Watkins, E. J., Almhjell, P. J. & Arnold, F. H. Direct Enzymatic Synthesis of a Deep-Blue Fluorescent Noncanonical Amino Acid from Azulene and Serine. *ChemBioChem* **21**, 80–

83 (2020).

26. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved Synthesis of 4-Cyanotryptophan and Other Tryptophan Analogues in Aqueous Solvent Using Variants of TrpB from Thermotoga maritima. *J. Org. Chem.* **83**, 7447–7452 (2018).

27. Buller, A. R. *et al.* Directed Evolution Mimics Allosteric Activation by Stepwise Tuning of the Conformational Ensemble. *J. Am. Chem. Soc.* **140**, 7256–7266 (2018).

28. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8**, 1–10 (2017).

29. Dick, M., Sarai, N. S., Martynowycz, M. W., Gonen, T. & Arnold, F. H. Tailoring Tryptophan Synthase TrpB for Selective Quaternary Carbon Bond Formation. *J. Am. Chem. Soc.* **141**, 19817–19822 (2019).

30. Romney, D. K., Sarai, N. S. & Arnold, F. H. Nitroalkanes as Versatile Nucleophiles for Enzymatic Synthesis of Noncanonical Amino Acids. *ACS Catal.* **9,** *8726-8730* (2019).

31. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.* **4**, 617–623 (2008).

32. Makałowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 9407–9412 (1998).

33. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, (2011).

34. Stiffler, M. A. *et al.* Protein Structure from Experimental Evolution. *Cell Syst.* **10**, 15-24.e5 (2020).

35. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

36. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS

carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

37. Ryan, O. W. & Cate, J. H. D. Multiplex engineering of industrial yeast genomes using CRISPRm. *Methods in Enzymology* **546**, 473-489 (Elsevier Inc., 2014).

38. Jung, P. P., Christian, N., Kay, D. P., Skupin, A. & Linster, C. L. Protocols and programs for high-throughput growth and aging phenotyping in yeast. *PLoS One* **10**, (2015).

39. Lane, A. N. & Kirschner, K. The Catalytic Mechanism of Tryptophan Synthase from Escherichia coli. *Eur. J. Biochem.* **129**, 571–582 (1983).

40. Goodsell, D. S., Autin, L. & Olson, A. J. Illustrate: Software for Biomolecular Illustration. *Structure* **27**, 1716-1720.e1 (2019).

41. Eigen, M., McCaskill, J. & Schuster, P. Molecular quasi-species. *J. Phys. Chem.* **92**, 6881–6891 (1988).

## 2.7 Author List

Gordon Rix, Ella J. Watkins-Dulaney, Patrick J. Almhjell, Christina E. Boville, Frances H. Arnold, and Chang C. Liu

## 2.8 Author Contributions

All authors contributed to experimental design and data analysis. G.R. cloned all genetic constructs; set up, performed, and characterized evolution experiments; and carried out yeast growth rate experiments. E.J.W. performed the panel HPLC-MS assay and indole conversion rate measurements on *Tm*TrpBs from variant set 2, and P.J.A. analyzed the results. C.E.B. performed *in vitro* characterizations of *Tm*TrpBs from variant set 1 and performed the thermal shift assay and substrate scope characterizations for *Tm*TrpB variants WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A. P.J.A. performed enzyme kinetics assays for *Tm*TrpB variants WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A. G.R. and C.C.L. wrote the manuscript with input and contributions from all authors. Authors: Gordon Rix, Ella J. Watkins-Dulaney, Patrick J. Almhjell, Christina E. Boville, Frances H. Arnold, and Chang C. Liu.

# Chapter 3. Continuous evolution of user-defined genes at 1-million-times the genomic mutation rate

## 3.1 Abstract

When nature maintains or evolves a gene's function over millions of years at scale, it produces a diversity of homologous sequences whose patterns of conservation and change contain rich structural, functional, and historical information about the gene. However, natural gene diversity likely excludes vast regions of functional sequence space and includes phylogenetic and evolutionary eccentricities, limiting what information we can extract. We introduce an accessible experimental approach for compressing long-term gene evolution to laboratory timescales, allowing for the direct observation of extensive adaptation and divergence followed by inference of structural, functional, and environmental constraints for any selectable gene. To enable this approach, we developed a new orthogonal DNA replication (OrthoRep) system that durably hypermutates chosen genes at a rate of >$10^{-4}$ substitutions per base *in vivo*. When OrthoRep was used to evolve a conditionally essential maladapted enzyme, we obtained thousands of unique multi-mutation sequences with many pairs >60 amino acids apart (>15% divergence), revealing known and new factors influencing enzyme adaptation. The fitness of evolved sequences was not predictable by advanced machine learning models trained on natural variation. We suggest that OrthoRep supports the prospective and systematic discovery of constraints shaping gene evolution, uncovering of new regions in fitness landscapes, and general applications in biomolecular engineering.

## 3.2 Introduction

Over the history of life, evolution has carried out a large-scale experiment exploring how gene sequences change under the constraints of prevailing or shifting structural and functional demands. The results of this natural experiment, embedded within the patterns of diversity across extant gene sequences, are of fundamental value to almost all areas of life sciences. For example, sequence conservation within a gene family is used to identify functionally critical residues [1–4], covariation among positions in an RNA or protein is used to deduce structural contacts and sectors of connectivity [5–12], differences in amino acid composition reveal environmental preferences (*e.g.*, temperature [13–15] or subcellular localization [16]), and differences in the conserved physicochemical properties across regions of a protein reflect driving forces behind folding [17,18]. Natural diversity across homologs also serves as a shared biomolecular engineering resource that can be mined for desired activities or recombined to access new functions [19–25]. Additionally, machine learning (ML) models have proven incredibly effective at extracting meaningful representations of biomolecular structure and function from the extensive diversity within and across gene families, as exemplified by their ability to predict functional effects of mutations [26–28], design functional sequences [29–31], and predict protein structures [32–34]. However, the natural evolution of highly diverse gene sequences under the constraints of selective forces — or conversely, the imprinting of selective forces and design principles into the statistics of sequence diversity — takes a long time at the slow rates of mutation in cellular and multicellular organisms. For example, reaching the 11% median divergence separating essential mouse and human genes [35] took ~96 million years [36]. Moreover, generating extensive collections of diverged sequences required complex histories of geographical isolation and speciation that allowed many populations to evolve separately to maintain variation in the face of within-population selective sweeps or genetic drift. Is it possible to compress long and vast gene evolution processes into laboratory experiments? Doing so would allow us to systematically detect novel structural and functional constraints governing biology, engineer custom biomolecules, create rich new sources of genetic variation for neofunctionalization or ML, and prospectively study the mechanisms and principles by which histories of selective forces become embedded into the patterns of sequence diversity.

We and others have endeavored towards this goal [37,38] through the development of scalable accelerated continuous evolution systems [39,40], such as our orthogonal DNA replication (OrthoRep) system in *Saccharomyces cerevisiae* [41,42]. OrthoRep cells have an additional DNA replication system comprising an orthogonal DNA polymerase (DNAP)/plasmid pair wherein the orthogonal DNAP (TP-DNAP1) durably replicates the orthogonal plasmid (p1) but does not replicate the host genome; likewise, host DNAPs replicate the host genome but not p1 (78Fig. 3.1A). Through this architecture, OrthoRep supports the sustainable coexistence of two independent mutation rates in the same cell: a low mutation rate of $10^{-10}$ substitutions per base (s.p.b.) for the large host genome, and a high mutation rate of $10^{-5}$ s.p.b. exclusively acting on p1 encoding only user-defined genes. OrthoRep's $10^{-5}$ s.p.b. mutation rate exceeds the error thresholds of the host genome [42], allowing us to drive the rapid, continuous evolution of chosen genes as cells autonomously propagate. However, $10^{-5}$ s.p.b. is not high enough to observe extensive evolution on laboratory timescales in the general case where evolution occurs both with and without positive selection. In the specific case of evolution under strong positive selection, sufficiently large population sizes can directly compensate for moderate mutation rates by increasing the beneficial mutation supply on which selection "pulls"; in this case, the OrthoRep system has successfully evolved enzymes [38,43,44], biosynthetic pathways [45], biosensors [46], drug targets [42], and antibodies [47,48] through long adaptive mutational pathways. Yet in the general case that includes when purifying selection is dominant or when selection is absent, both highly relevant in the generation of natural diversity and the ability to escape local fitness optima, mutation becomes the main force pushing sequence change. Without the pull of positive selection, our previous OrthoRep systems would take 100 generations (8-12 days for the yeast host of OrthoRep) just to sample an average of 1 new mutation in a typical 1 kb gene.

Here, we present the engineering of novel OrthoRep systems that have a mutation rate up to 1.7 × $10^{-4}$ s.p.b., corresponding to a new mutation in a typical 1 kb gene once every <10 generations in the absence of any selection. This intensified mutational force on chosen genes, operating at

>1 million times the mutation rate of the host genome, allows us to mimic extended periods of natural gene evolution on laboratory timescales in the general case. We describe the TP-DNAP1 engineering effort leading to these novel OrthoRep systems and provide characterization of their performance through detailed measurements of their elevated mutation rates, reduced mutational biases, and durability. We then show how, in <3 months of laboratory passaging (totaling <15 hours of researcher intervention) of 96 independent populations, a conditionally essential gene encoded on p1 diverges to an extent where the median distance separating pairs of evolved sequences is 35 amino acids, with thousands of unique pairs separated by >60 amino acids. This corresponds to an amino acid divergence of ~9% to >15%, exceeding the median 11% distance separating orthologous genes between mouse and human [35]. By analyzing the rich collection of diverged sequences throughout their laboratory evolutionary history, we uncover hidden forces shaping and constraining sequence change, such as a preference for negative net charge, supporting a proposed mechanism by which proteins avoid large-scale indiscriminate clustering in the crowded environment inside cells [49]. We also extract examples of allosteric network remodeling through the cooccurrence of mutations across distinct clades and temperature optimization through amino acid content change. Overall, our work provides an approach to systematically reveal the evolutionary constraints and selective forces that genes experience and delivers an upgraded OrthoRep system for broad application.

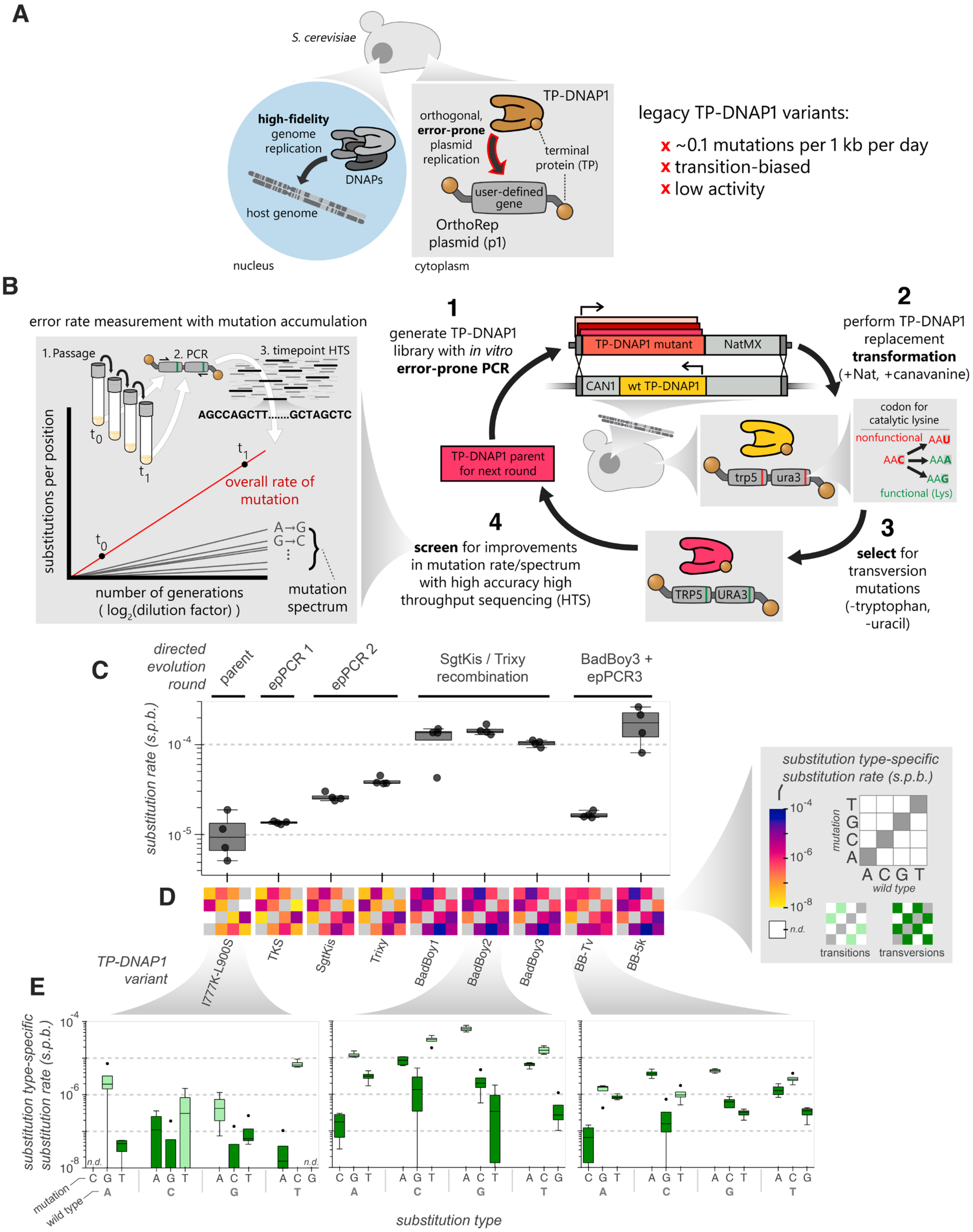


**Fig. 3.1. Engineering orthogonal DNA polymerases for increased mutation rates.** (**A**) Architecture of the OrthoRep system. A DNA polymerase (TP-DNAP1) that exclusively replicates a specific cytoplasmically

localized plasmid via protein primed replication at a high error rate enables *in vivo* targeted mutagenesis without mutagenizing genomic DNA. (**B**) Schematic for a directed evolution approach to improving TP-DNAP1's mutation rates and mutation spectrum incorporating both a direct selection for rare transversion mutations as well as high accuracy mutation rate measurement using a mutation accumulation and high throughput sequencing (HTS) assay. (**C**-**E**) Mutation rate measurements via mutation accumulation for a series of TP-DNAP1 directed evolution intermediates showing either overall mutation rates as boxplots (C), mean mutation rates for individual substitution types as heatmaps (D), or mutation rates for individual substitution types for three individual TP-DNAP1 variants as boxplots (E). Points are representative of individual biological replicates, each representing 3-4 timepoints with >50 sequences each. Boxplots and heatmaps are representative of n=4 biological replicates. Box plot central line, boxes, and whiskers represent the median, interquartile range, and minimum / maximum values, respectively.

## 3.3 Results

### Motivation and preparation for OrthoRep engineering

The current state-of-the-art OrthoRep system uses TP-DNAP1-4-2 as the error-prone orthogonal DNAP [42]. Besides its suboptimal error rate of $10^{-5}$ s.p.b., TP-DNAP1-4-2 also has low replicative activity (Fig. S3.1) and exhibits a heavily transition-biased mutation spectrum (Fig. S3.2) [42], suppressing the impact of point mutations on amino acid sequence during protein evolution (Fig. S3.3). To increase OrthoRep's overall error rate, transversion rate, and activity, we carried out a directed evolution campaign on TP-DNAP1.

To prepare the directed evolution campaign, we engineered a genetic selection strain, OR-Y488, that could enrich TP-DNAP1s with increased error rates and activities from libraries of TP-DNAP1 variants (Fig. 3.1B). We prioritized selection for increased transversion rates under the rationale that TP-DNAP1s with high transversion rates would also have high overall error rates, because transversions require tolerance of larger structural aberrations than transitions [50]. OR-Y488 contained a p1 plasmid (p1-ura3*-trp5*) encoding two auxotrophic marker genes, *ura3* and *trp5*, each specifically disabled via an active site missense mutation whose sole option for functional reversion is a transversion (Fig. 3.1B and Fig. S3.4). TP-DNAP1s with the highest transversion rates should restore *URA3* and *TRP5* most frequently, resulting in their enrichment when OR-Y488 is grown in the absence of exogenous uracil or tryptophan. This selection was designed to occur in two sequential stages, first for *URA3* restoration and then for *TRP5* restoration, to suppress the enrichment of low error rate TP-DNAP1 variants in revertants that stochastically emerge from the long tail of the Luria-Delbrück distribution [51]. We also designed a counterselection-based high-efficiency integration strategy for transforming TP-DNAP1 variants into OR-Y488 such that only the variant TP-DNAP1 would replicate p1-ura3*-trp5* in transformed cells (see Methods and Fig. S3.5).

To precisely guide the TP-DNAP1 directed evolution campaign, we developed a mutation accumulation assay [52] for p1 and coupled it with high throughput sequencing (HTS) of p1 amplicons using the Oxford Nanopore Technologies (ONT) platform [53], allowing us to accurately determine the rate for any individual type of mutation (Fig. 3.1B). In this assay, a strain containing p1 replicated by any TP-DNAP1 variant is grown for a set number of generations. A region of p1 not under selection is then amplified, tagged with unique molecular identifiers (UMIs) for sequencing error correction [54,55], and sequenced at two or more timepoints during mutation accumulation (see Methods). The rate of change in the number of mutations per position is calculated for all types of mutations to fully describe the overall mutation rate and mutation preferences of the TP-DNAP1 variant. To facilitate rapid characterization, we developed a custom analysis pipeline that could carry out most of the analysis steps autonomously (Fig. S3.6). This pipeline, Mutation Analysis for Parallel Laboratory Evolution or Maple, performs consensus sequence generation, demultiplexing, mutation identification, mutation rate analysis, and many other operations to generate a collection of visualizations and data tables that accelerate analysis of mutation-rich sequencing datasets while minimizing user input.

## Directed evolution of a high error rate orthogonal DNAP

With the genetic selection, HTS-based mutation accumulation measurement pipeline, and Maple in place, we carried out our TP-DNAP1 directed evolution campaign over five rounds, including three rounds of error-prone PCR (epPCR) and selection (Table S3.1). In round 1, we started from an epPCR library generated from TP-DNAP1(I777K, L900S). TP-DNAP1(I777K, L900S) is a relative of TP-DNAP1-4-2 with a mutation rate near $10^{-5}$ s.p.b. and greater activity than TP-DNAP1-4-2 [42] (Fig. S3.1). We reasoned that its higher activity would confer mutational robustness, increasing the fraction of active library members when used as the parental sequence. After integration of the epPCR library into OR-Y488 and enrichment of TP-DNAP1s that could restore *URA3* via a transversion mutation (Fig. S3.4C), we isolated ~90 colonies and screened for their ability to restore *TRP5* through the second selection stage, which was done in multiple replicates per clone following a fluctuation analysis forma [42] to obtain estimates on phenotypic mutation rate. We isolated several clones whose phenotypic mutation rates were up to 10-fold elevated (Figs. S7A) and evaluated their genotypic mutation rates in detail using our mutation accumulation assay and Maple. Among these clones, TP-DNAP1-TKS, with the mutation P680T, had the highest per base mutation rate (Fig. S3.7B-C). In round 2, selection applied to an epPCR library generated from TP-DNAP1-TKS resulted in the enrichment of several clones whose full mutation rates and spectra were then determined (Fig. S3.8). These several clones turned out to represent two unique TP-DNAP1 variants. The two variants satisfied the requirements of selection via distinct strategies. One variant, TP-DNAP1-SgtKis, contained 5 nonsynonymous mutations in addition to those in the parent TP-DNAP1-TKS and demonstrated an altered mutation spectrum favoring transversions but only a minimal apparent increase in overall mutation rate. Another variant, TP-DNAP1-Trixy contained three nonsynonymous mutations and had the highest overall mutation rate measured, but only marginal changes to the mutation spectrum.

Since TP-DNAP1-SgtKis had an increased transversion rate and TP-DNAP1-Trixy had an increased overall substitution rate, we reasoned that their combination could yield orthogonal DNAPs with both high overall and transversion rates. We cloned seven new TP-DNAP1 variants where a subset of mutations from TP-DNAP1-SgtKis were added to TP-DNAP1-Trixy and obtained their fully described mutation rates (SgtKis / Trixy recombination round, Fig. S3.9). Remarkably, all TP-DNAP1 variants that included the mutations L474S and E488G from TP-DNAP1-SgtKis exhibited a dramatic elevation in their overall mutation rate, in each case bringing the per base rate to ~$10^{-4}$ s.p.b. (Fig. 3.1C, Fig. S3.9C), 1-million-fold higher than the yeast genomic mutation rate [42,56]. Furthermore, the broad mutation spectrum of TP-DNAP1-SgtKis was preserved in these variants (Fig. 3.1D-E, Fig. S3.9D), which we named BadBoy1, BadBoy2, and BadBoy3 (Table S3.1) to recognize their poor fidelity.

Our directed evolution campaign yielded two additional notable TP-DNAP1 variants resulting from combining mutations enriched from an epPCR library derived from TP-DNAP1-Trixy (epPCR 3 round, Fig. S3.10) with those in BadBoy3 (BadBoy3 + epPCR 3 round, Table S3.1). One of the resulting TP-DNAP1s, named BB-5k, exhibited a further increase in mutation rate, to $1.7 \times 10^{-4}$ s.p.b. (Fig. 3.1C), representing ~1 mutation every time a 5 kb recombinant p1 plasmid is replicated. However, BB-5k did not durably maintain p1-ura3*-trp5* in two of four biological replicates over the ~120 generations of mutation accumulation tested, possibly because it exerts an excessive mutational load on the *LEU2* marker used to maintain p1-ura3*-trp5*. The other DNAP of potential value, BB-Tv, has a relatively low mutation rate ($1.6 \times 10^{-5}$ s.p.b.), like that of our previous OrthoRep systems. Yet unlike previous systems, BB-Tv demonstrated a near-ideal mutation spectrum, with transversions accounting for 43% of all mutations, up from only 2.5% for TP-DNAP1(I777K, L9000S) (Table S3.1, Fig. 3.1D). BB-Tv should therefore be useful in continuous evolution experiments involving larger targets that have lower error thresholds.

Detailed characterization of mutation accumulation across our several TP-DNAP1 variants revealed interesting trends. For example, when we compared mutation rates across different regions of p1, we found that mutation frequencies were consistently high in regions of p1 not under selection but were lower in regions encoding genes under selection (Fig. S3.11A-B), an effect that was most pronounced for the highest mutation rate TP-DNAP1 variants ($R^2 \approx 0.5$, Fig. S3.11C). This implies that the $10^{-4}$ s.p.b. mutation rates of our TP-DNAP1s have entered a regime in which, without purifying selection, the function of a gene is quickly degraded, suggestive of nearing gene error thresholds. We also found that the mutation rate of TP-DNAP1-Trixy was correlated with p1 length while TP-DNAP1-SgtKis did not exhibit this trend, suggesting an interplay between mutation rate and the number of bases replicated that is dependent on mutation mechanism (Fig. S3.12). For BadBoy1, BadBoy2, and BadBoy3, mutation rates were largely independent of p1 length (Fig. S3.12). Finally, examination of substitution-type-specific mutation rates among our TP-DNAP1 variants showed that mutation of position 777 to either Ser or Thr yields a large drop (~10-fold) in the A:T→G:C transition mutation rate while having little effect on the reverse G:C→A:T rate (Fig. S3.13), demonstrating that even similar mutation types can be generated by independent mechanisms. These subtle observations that come through our precise and rigorous mutation rate measurement pipeline should aid the future engineering of OrthoRep and other continuous evolution systems [57].

Overall, our orthogonal TP-DNAP1 directed evolution campaign, in conjunction with past efforts, complete a set of OrthoRep systems evenly spanning a range of ~5 orders-of-magnitude, from ~$10^{-9}$ s.p.b., similar to the mutation rate of modern cellular genomes, up to ~$10^{-4}$ s.p.b., far beyond the error thresholds of modern cellular genomes [42] and likely approaching the error threshold of individual genes where maximal adaptation rates can be reached [58,59].

## Extensive divergence of a conditionally essential gene on laboratory timescales

Our new OrthoRep systems should be capable of driving rapid evolution of chosen genes regardless of the type of selection imposed. We encoded the β-subunit of *Thermotoga maritima's* tryptophan synthase (TrpB) onto p1 in a tryptophan auxotroph where p1 is exclusively replicated by BadBoy2 at a mutation rate of 1.4 × $10^{-4}$ s.p.b. (Fig. 3.2A). TrpB condenses indole with serine to yield tryptophan (Trp), but *T. maritima* TrpB is maladapted for this standalone reaction, since it normally functions in complex with TrpA [60,61]. Therefore, cells grown in the absence of Trp need to evolve improved TrpB activity to propagate, allowing TrpB to serve as the subject of an extended evolution experiment that included no selection, positive selection, and purifying selection phases (Fig. S3.14).

We designed our evolution experiment to prioritize sequence divergence and diversity in order to maximize the amount of evolutionary information that could later be extracted. The evolution experiment was therefore run for ~540 generations (<3 months) at the scale of 96 independent replicate 500 µL cultures. 1:1024 (10 generation) transfers into fresh growth medium were made every one or two days, depending on cell density, following a passaging schedule that included all types of selection pressures: 'no selection' phases, 'positive selection' phases, and 'purifying selection' phases, as summarized in Fig. S3.14. We collected DNA from cells at 15 timepoints throughout the ~540-generation evolution experiment and used a high-yield rolling circle amplification-based sequencing strategy in conjunction with HTS and Maple to analyze the TrpB sequences sampled from these timepoints (See Methods and Fig. S3.15).

Overall, we observed a wide diversity of evolutionary outcomes (Fig. S3.16) and a monotonic increase in both the average number of mutations and diversity (as measured by pairwise hamming distances) in TrpB throughout all phases of evolution (Fig. 3.2B-D). The rate at which mutations accumulated in the population was highest in the no selection phase (~0.15 amino acid changes per generation) followed by the positive selection and purifying selection phases, which exhibited similar rates (~0.024 and ~0.018 amino acid changes per generation, respectively) (Table S3.2). Notably, the positive selection phases did not have the highest rate of mutation accumulation even though there was substantial adaptation in producing TrpBs that supported cell growth in the absence of Trp and presence of moderate concentrations of indole. This suggests that BadBoy2's error rate was high enough to consistently "saturate" positive selection with an overabundance of beneficial mutations in TrpB, predicting the broader power of upgraded OrthoRep mutation rates in biomolecular evolution applications. Notably, mutation accumulation in the purifying selection phases was appreciable yet substantially slower than in the no selection and positive selection phases. From this, we conclude that BadBoy2's error rate was high enough to constantly test the constraints of structure and function, demonstrating the general utility of OrthoRep in uncovering biological forces governing how genes and biomolecules operate. At the end of the evolution experiment, each TrpB sequence had an average of 20.6 amino acid and 44.5 nucleotide mutations (Table S3.2). In the last two timepoints, over 1800 sequences (~5%) had accumulated more than 30 amino acid changes from the ancestral 398 amino acid wt TrpB. The distribution of pairwise hamming distances showed that sequences had substantially diverged from each other (Fig. 3.2C-D) such that in the final timepoint, 24% of sequences were separated from each other by 40 amino acids or more, including more than 4000 sequence pairs differing by a pairwise amino acid hamming distance of at least 60 (Fig. 3.2D, inset). This level of sequence divergence (>15%) approximates that between human and mouse essential gene orthologs [35]. Indeed, our experiment demonstrates that we can compress extensive and complex gene evolution processes into laboratory timescales, resulting in the generation of highly diverse sequence families shaped by varied selection conditions over long mutational pathways. What does the evolutionary information embedded into this diversity reveal?

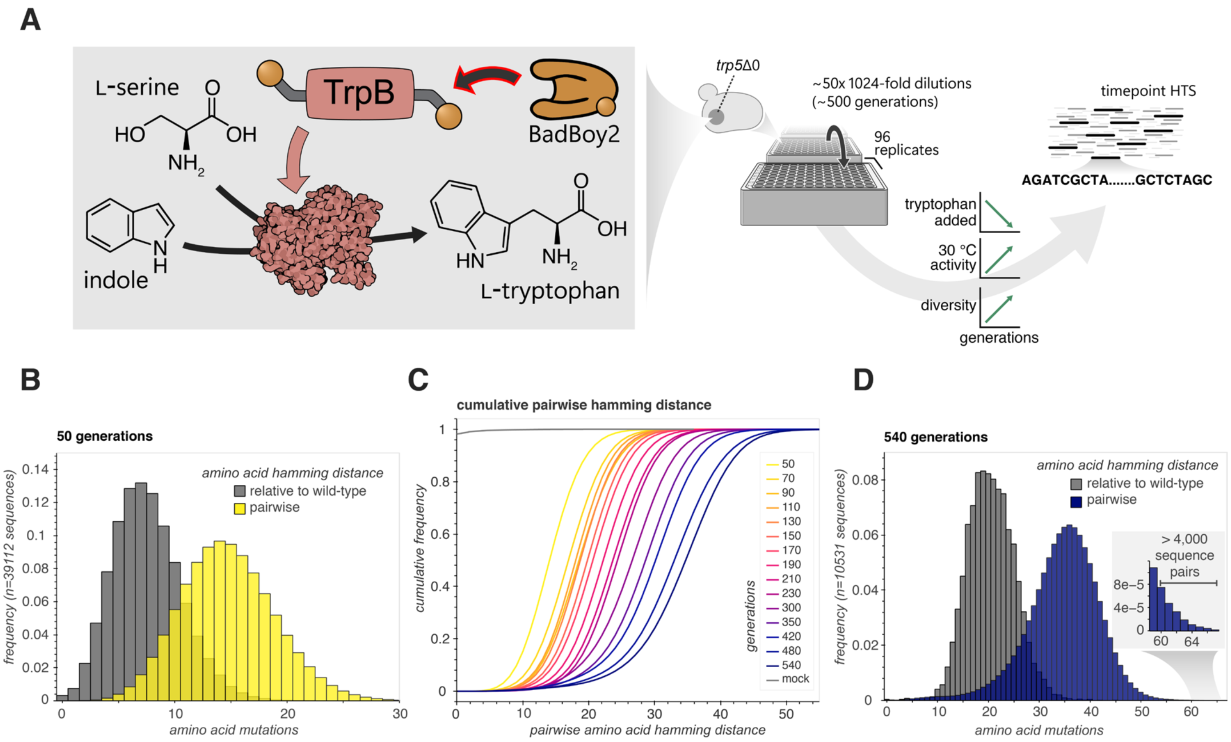


**Fig. 3.2. Massively parallel diversification and evolution of TrpB.** (**A**) Schematic for OrthoRep evolution of the tryptophan synthase β-subunit from *Thermotoga maritima* (TrpB) for standalone function in yeast. TrpB was integrated onto the p1 plasmid in a yeast strain lacking the native yeast tryptophan synthase gene (*TRP5*). 96 independent cultures of the resulting strain were passaged mostly under selective pressure for Trp production using exogenously supplied indole over ~500 generations. DNA from fifteen timepoints throughout the evolution campaign was harvested and sequenced using HTS. (**B**-**D**) Pairwise amino acid hamming distances as distributions for both pairwise and relative to the wild-type sequence at the first and last timepoint (B and D respectively), and as pairwise cumulative distributions for all timepoints (C).

## General structural and functional constraints

~500,000 sequences of TrpB with an average of 13.1 amino acid replacements each were captured over the evolution experiment, and over 90% of those sequences were unique (Table S3.2). With such a diverse evolutionary dataset, patterns of conservation should contain structural and functional constraints defining TrpB. To test this notion, we used an AlphaFold structure [33] and knowledge from previous studies on TrpB [62,63] to first categorize each residue in TrpB according to its general structural or functional role, as outlined in Fig. 3.3A. We then asked whether different categories showed different levels of conservation. We immediately noticed a congruence between relative conservation and buried residues, revealing the well-known importance of a buried hydrophobic core in protein folding (Fig. 3.3B). We also noticed that residues within 5 Å of TrpB's active site were highly conserved. Additionally, there was a relative abundance of amino acid replacements at certain positions in the COMM domain, suggesting that it was a target of adaptation.

To improve our resolution in such observations, we generated a simulated dataset of TrpB mutants where each sequence accumulated mutations from the exact encoding of wt TrpB using BadBoy2's fully described mutation biases. For each simulated sequence, mutation accumulation was stopped once it matched the number of synonymous mutations of a corresponding sequence from the real dataset. The simulated dataset serves as the null model where patterns in evolved TrpB sequences are simply a reflection of the mutation preferences of BadBoy2 and codon usage of wt TrpB. Under the assumption that synonymous mutations have no effect on fitness, the nonsynonymous differences between the real dataset and the simulated dataset contain the influence of selective forces. An excess of nonsynonymous changes in the real dataset compared to the simulated dataset is therefore an indication of positive selection, while the opposite signifies purifying selection. As shown in Fig. 3.3C, at the generation 70 timepoint, the real dataset has a paucity of nonsynonymous mutations per sequence in the active site region and buried residues and an excess of nonsynonymous mutations per sequence in the COMM domain. Generation 70 is during the first phase of positive selection for TrpB's operation as a standalone enzyme capable of generating tryptophan (Fig. S3.14), so this timepoint is most likely to reveal signatures of adaptation. (This is supported by Fig. S3.17, which shows the overall mutation accumulation dynamics; generation 70 is the timepoint with the greatest excess of nonsynonymous mutations relative to the simulated dataset among the timepoints for which selection for TrpB function had been applied.) At generation 70, we find that positive selection had enriched mutations to buried and COMM domain residues while purifying selection had already removed changes to the active site region. Our detection of the COMM domain as a focus of adaptation can be explained, since

TrpB is a well-studied enzyme. The COMM domain mediates allosteric activation of TrpB by TrpA [60]. As our evolution experiment required TrpB to operate in a standalone manner without TrpA, the remodeling of allosteric networks through the COMM domain was an expected means to adaptation in line with previous studies of engineered TrpB standalone activity [62,63]. Had TrpB not been well-studied, the detection of this critical region for adaptation from the evolutionary information could have suggested such an explanation.

We also considered the final timepoint of the evolution experiment. As shown in Fig. 3.3C and Fig. S3.17, by generation 500, the influence of purifying selection had dominated, as evidenced by the paucity of mutations in the real data compared to the simulated. Although purifying selection constrained all regions of TrpB, some were clearly more constrained than others. For example, the active site region had almost no mutations (maximally 1 or 2 but mostly 0) and deviated from the simulated mutant distribution more than all other regions. Buried residues also had substantially fewer nonsynonymous mutations than the simulated dataset. In contrast, the effect of purifying selection was less pronounced on surface residues, reflecting the relative tolerance of protein surfaces to mutation. Surprisingly, this also applied to the newly solvent-exposed β-α interface region. In the absence of the α-subunit, this region should be more solvent-exposed than in TrpB's native context. The fact that there was little noticeable difference in the effects of selection on the β-α and β-β interfaces suggests that solvent-exposure of this region had minimal impact on TrpB fitness.

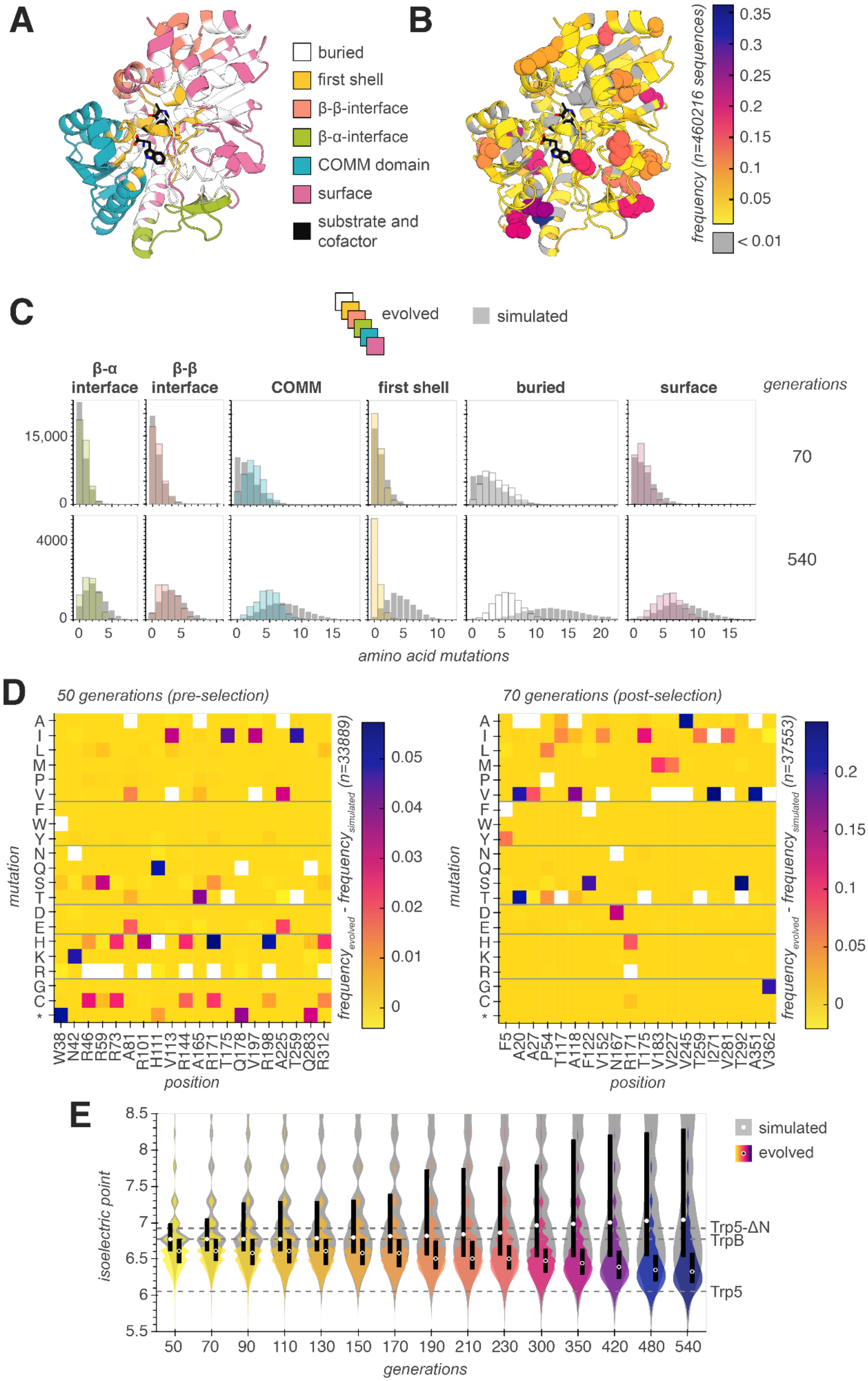

A
buried
first shell
β-β-interface
β-α-interface
COMM domain
surface
substrate and cofactor
B
frequency (n=460216 sequences)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
< 0.01
C
evolved
simulated
β-α interface
β-β interface
COMM
first shell
buried
surface
generations
15,000
0
70
4000
0
540
0 5
0 5
0 5 10 15
0 5 10
0 5 10 15 20
0 5 10 15
amino acid mutations
D
50 generations (pre-selection)
70 generations (post-selection)
A
I
L
M
P
V
F
W
Y
N
Q
S
T
D
E
H
K
R
G
C
*
mutation
W38 N42 R46 R59 R73 A81 R101 H111 V113 R144 A165 R171 T175 Q178 V197 R198 A225 T259 Q283 R312
position
frequency_evolved - frequency_simulated (n=33889)
0.05
0.04
0.03
0.02
0.01
0
F5 A20 A27 P54 T117 A118 F122 V152 N167 R171 T175 V183 V227 V245 T259 I271 V281 T292 A351 V362
position
frequency_evolved - frequency_simulated (n=37553)
0.2
0.15
0.1
0.05
0
E
isoelectric point
8.5
8
7.5
7
6.5
6
5.5
simulated
evolved
Trp5-ΔN
TrpB
Trp5
50 70 90 110 130 150 170 190 210 230 300 350 420 480 540
generations

**Fig. 3.3. Revealed effects of selective pressures.** (**A**) AlphaFold structure of *Thermotoga maritima* TrpB with different regions colored according to their structural role. First shell, β-β interface, and β-α interface residues are designated as such if they are within 5 Å of the substrate and cofactor (Trp and PLP), the other β subunit, or the other α subunit in the αββα heterotetramer holoenzyme, respectively. Alignment to *Pyrococcus furiosus* TrpB crystal structures (PDB codes 5E0K and 5DW3) were used to determine distances from substrate and cofactor, α-subunit, and β-subunit. Mean solvent accessible surface area (SASA) was used to categorize all remaining residues as either surface (SASA≥0.2) or buried (SASA<0.2). (**B**) Heatmap of mutations among OrthoRep-evolved TrpB sequences applied to the AlphaFold structure. (**C**) Distributions of mutations among OrthoRep-evolved TrpB sequences within the 6 structural regions compared to a simulated dataset of sequences with the same number of synonymous mutations. (**D**) Heatmap of mutation frequency for all mutations among the 20 most frequently mutated positions in the timepoint corresponding to either 50 or 70 generations. Frequencies for simulated sequences are subtracted to account for bias due to BadBoy2's mutation preference and wt TrpB sequence content. Isoelectric points of the wild type TrpB, the TrpA-TrpB holoenzyme ortholog from *Saccharomyces cerevisiae* (Trp5) and an N-terminally-truncated Trp5 homologous to TrpB (Trp5-ΔN) are shown for comparison. (**E**) Violin plots of isoelectric points for all OrthoRep-evolved and simulated TrpB sequences, split by timepoint. Points and black bars denote the means and interquartile range for all sequences within each timepoint.

### Isoelectric point evolution for intracellular compatibility

We examined the 20 residues that were mutated in greatest excess among real evolved sequences relative to simulated sequences in the first two timepoints (Fig. 3.3D). Despite the entire wt TrpB protein containing only 19 arginine residues in total (5%), arginine constituted 8 of these 20 most frequently mutated residues by generation 50. This came as a surprise because this enrichment occurred even before selection for TrpB function was imposed. This led us to hypothesize charge optimization as a driving selective force, because charge could influence not only TrpB function itself, but also the cellular environment within which TrpB operated. To evaluate this hypothesis, we calculated the isoelectric point (pI) of sequences throughout the evolution experiment and examined its distribution over time (Fig. 3.3E). Indeed, we found that the pI of sequences was significantly lower at the end of the experiment than early in the experiment ($p<0.0001$, Mann-Whitney U test). Comparison of pI change with simulation corroborates that this effect was driven by positive selection. A similar analysis of hydrophobicity revealed a modest decrease in hydrophobicity over time for simulated sequences that was mitigated by selection in the real data (Fig. S3.18). The change in hydrophobicity for the real sequences throughout the experiment was less pronounced than the change in pI, however, suggesting that charge optimization, and not polarity in general, was the dominant selective force. Notably, the majority of the shift in the pI distribution occurred in the latter half of the experiment (Fig. 3.3E), highlighting the importance of sustained rapid mutagenesis over long periods of evolution to embed such presumably subtle selective forces into the data.

TrpB's pI evolved to be comfortably below the typical yeast cytosolic pH of 6.8 to 7.2 [64], which is consistent with the notion that intracellular proteins prefer to be negatively charged to minimize large-scale clustering with RNAs and other proteins [49,65]. One possible mechanism by which this preference could have driven the observed adaptation is through its influence on TrpB function itself, for example by increasing the diffusivity of the enzyme [66]. Another mechanism by which a preference for negative charge in TrpB could have been adaptive is by lessening its perturbation on other entities in the cell, for example by preventing spurious association or aggregation that would disturb the function of the proteome [49]. Our data does not exclude either mechanism but suggests that the latter mechanism is present. In generations 0-50 of the evolution experiment, TrpB was not under selection for function as excess Trp was supplied to the growth media. Indeed, HTS at the end of 50 generations detected the presence of many stop codons, which were mostly eliminated soon after positive selection for TrpB adaptation had been imposed (for example, HTS at the end of generation 70) (Fig. 3.3D). Yet at the end of 50 generations, pIs were significantly lower than that of the simulation ($p<0.0001$). Our observation of charge optimization even when there was no selection for the enzyme's function demonstrates that the intracellular context can impose constraints on the physicochemical properties of proteins independent of its primary molecular function. It also highlights the value of evolving proteins *in vivo* where subtle constraints dictating intracellular compatibility can both be revealed and included in the evolutionary optimization of protein function.

## Thermoadaptation

Given that our parental *T. maritima* TrpB was from a thermophile but needed to evolve standalone activity in a mesophile, we looked for statistical evidence of thermoadaptation in our evolved sequences. Haney *et al.* studied the patterns of amino acid replacements between natural orthologous proteins in mesophilic versus thermophilic organisms and found 17 amino acid replacements that distinguished the mesophilic variants from the thermophilic variants at homologous positions in multiple sequence alignments with high confidence [13]. When we evaluated the frequency of these 17 amino acid replacements among all mutations in our evolution experiment's outcomes, we found that replacements in the mesophilic direction were enriched (Fig. S3.19). As before, this illustrates the ability of extensive gene evolution to reveal selective forces through the evolutionary information embedded into the resulting diversity.

## Networks of coupled mutations

In addition to general selective forces, we investigated finer patterns in the outcomes of TrpB evolution with the expectation that these may reveal coevolving networks of amino acids responsible for adaptation. At the beginning of our evolution experiment, we had included short barcodes adjacent to the TrpB sequence integrated onto p1. This allowed us to 1) isolate the largest clades for analysis, since these should correspond to the fittest sequences, and 2) reduce the contribution of phylogeny by computationally limiting the number of sequences analyzed per clade (Fig. 3.4A). The latter increases the signature of mutations independently discovered across multiple clades, favoring the detection of mutations whose cooccurrence was functionally significant. Specifically, we considered clades whose barcode had at least 100 associated sequences — there were 93 such clades — and randomly downsampled to 100 sequences for clades whose members exceeded this number.

This analysis revealed that some sets of mutations frequently cooccur (Fig. 3.4B). In one particularly striking example, the A20V mutation was found to cooccur at a frequency above ~0.7 with a set of 5 other mutations (A118V, F122S, V245A, I271V, T292S) in 8 distinct lineages in later timepoints, implying a strong relationship among these mutations (Fig. 3.4C-D). This set of mutations, as well as other sets identified, are spread throughout the structure of TrpB, in line with recent studies demonstrating the prevalence of structurally distributed allosterically activating mutations in TrpB and other proteins [6,62,63,67]. Among these five mutations is the T292S mutation, which was previously identified in a TrpB directed evolution campaign as highly activating, alone conferring a >5-fold increase in the standalone catalytic efficiency of TrpB [20]. Intriguingly, the association among these five mutations was sensitive to their specific identities. For example, the A20T mutation was individually present at a higher frequency than A20V among all sequences (Fig. S3.16D) but was not associated with any specific other mutation at a frequency above ~0.4 (Fig. 3.4D). The overall picture from this analysis suggests the existence of heavily enriched individual mutations that are broadly activating across many sequence contexts and individual mutations that are only activating in combination with other mutations, implying long range epistatic interactions.

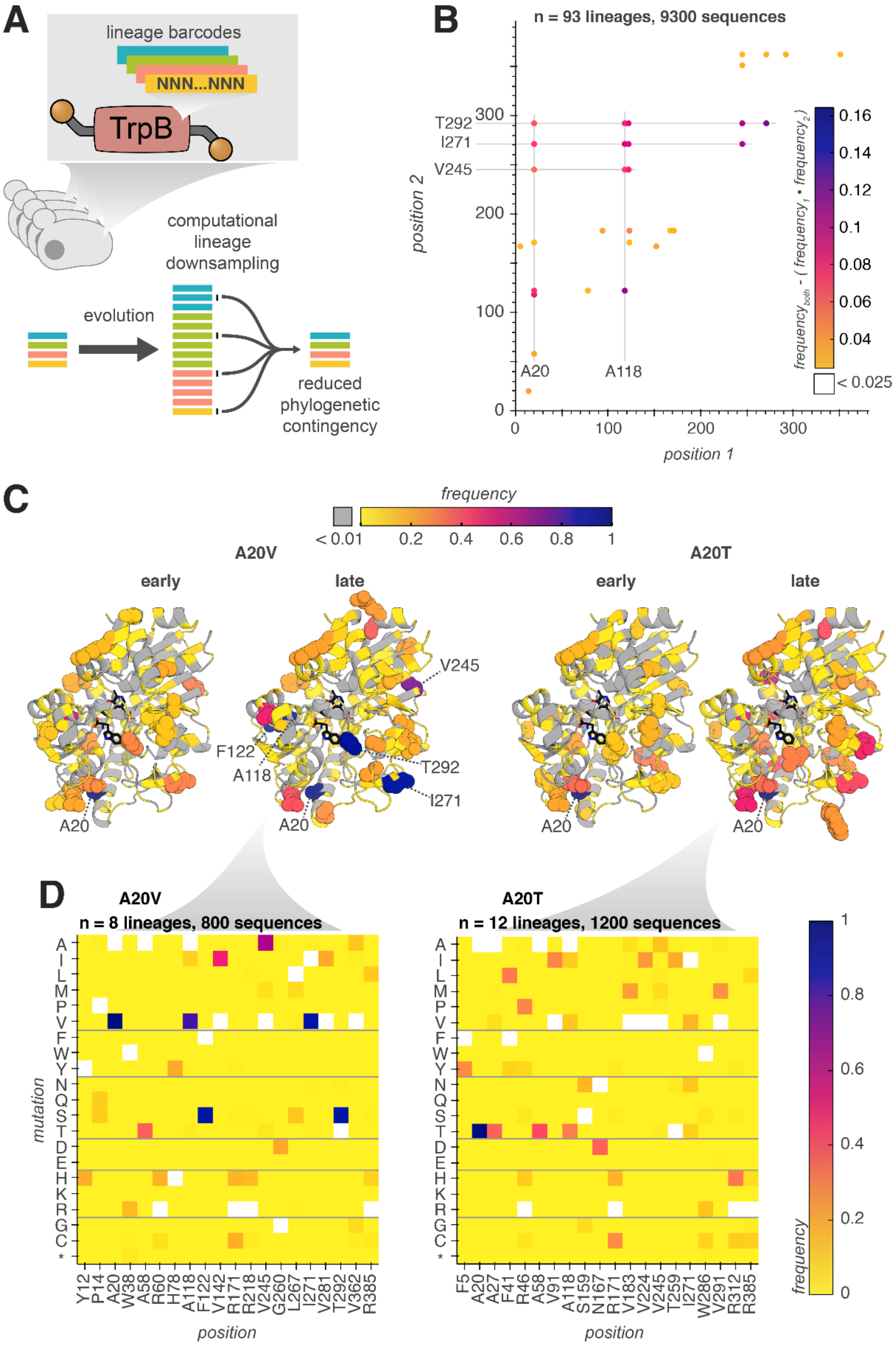

A
lineage barcodes
NNN...NNN
TrpB
computational lineage downsampling
evolution
reduced phylogenetic contingency
B
n = 93 lineages, 9300 sequences
T292
I271
V245
A20
A118
position 2
position 1
0
100
200
300
frequency$_{both}$ - ( frequency$_1$ • frequency$_2$ )
0.16
0.14
0.12
0.1
0.08
0.06
0.04
< 0.025
C
frequency
< 0.01
0.2
0.4
0.6
0.8
1
A20V
A20T
early
late
V245
F122
A118
T292
I271
A20
D
A20V
n = 8 lineages, 800 sequences
A20T
n = 12 lineages, 1200 sequences
mutation
A
I
L
M
P
V
F
W
Y
N
Q
S
T
D
E
H
K
R
G
C
*
Y12
P14
A20
W38
A58
R60
H78
A118
F122
V142
R171
R218
V245
G260
L267
I271
V281
T292
V362
R385
F5
A20
A27
F41
R46
A58
V91
A118
S159
N167
R171
V183
V224
V245
T259
I271
W286
V291
R312
R385
position
frequency
1
0.8
0.6
0.4
0.2
0

**Fig. 3.4. Lineage barcodes reveal covarying residues.** (**A**) Schematic of computational processing used to reduce phylogenetic contingency of residue covariation. (**B**) Residue covariation plot for the most frequently covarying residues among all timepoints in the TrpB evolution dataset for 93 lineages downsampled to 100 sequences per lineage. (**C**-**D**) Heatmaps of most frequently mutated residues among sequences containing mutations A20V or A20T, downsampled to 100 sequences and chosen from specific timepoints. Heatmaps of all positions for early (generations 70 and 90) and late (generations 480 and 540) timepoints are overlaid onto a TrpB AlphaFold structure with the 20 most frequently mutated positions shown as spheres (C) or shown as mutation-specific heatmaps of the most frequently mutated 20 positions in late timepoints (D).

### *Fitness of TrpBs and their predictability*

Accurately modeling the fitness landscapes of proteins is a major goal of ML. However, it is known that ML models are biased towards favoring sequences that are more similar to the natural sequences on which they were trained [68], and it remains unclear to what extent they can predict the function of sequences that are many mutations away from these natural sequences. It is also unclear whether ML can model how mutant sequences will perform on new functions that deviate from natural functions, in service of bioengineering goals such as enzyme and antibody engineering. To provide insight into these questions, we tested whether an advanced ML model could predict the fitness of our evolutionary outcomes.

To gain high-resolution fitness information on evolved TrpBs, we first profiled evolved variants in a high-throughput enrichment assay (Fig. 3.5A). We cloned a library of ~100,000 TrpB sequences isolated from our evolution experiment into a standard yeast plasmid that would not be subject to *in vivo* mutagenesis, transformed this library into yeast, applied selection for TrpB function, and tracked the enrichment or depletion of individual variants via HTS. Included in this library were two previously engineered control TrpB variants known to be either highly functional (TrpB-003-A) or nearly nonfunctional (*Tm*Triple) in the context of yeast Trp production complementation (Fig. 3.5B) [38]. We evaluated Trp production by members of this library using three distinct growth conditions: Trp-supplemented media (no selection), media lacking Trp with a high concentration of indole (400 μM, weak selection), and media lacking Trp with a low

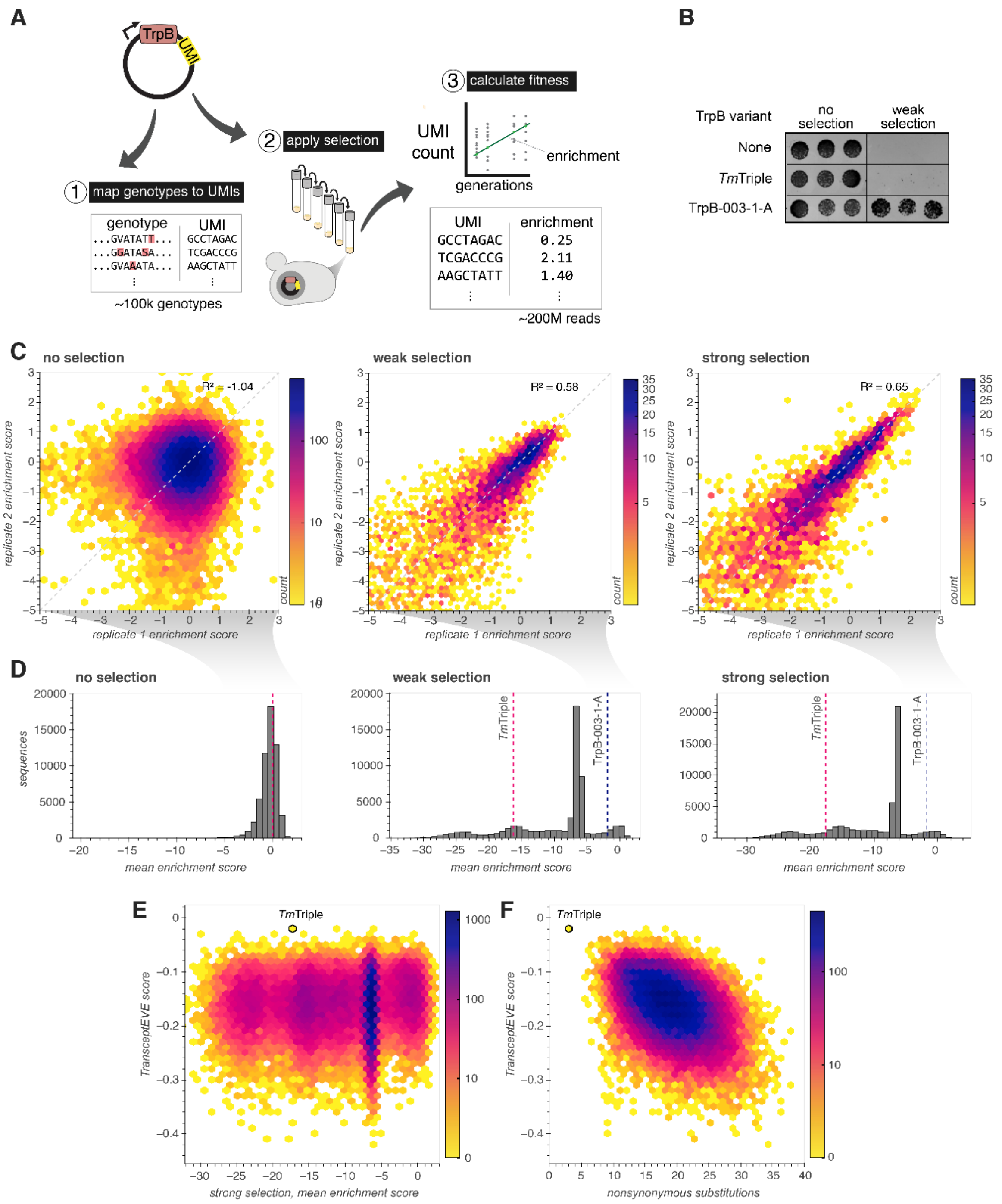


**Fig. 3.5. Pooled measurement and TransceptEVE prediction of TrpB variant fitness.** (**A**) Schematic of pooled TrpB fitness assay using HTS. (**B**) Spot plating growth assay of control sequences included in the pooled fitness assay. (**C**) Hexbin plot of replicate concordance among pairs of replicates under growth conditions with Trp (no selection), without Trp and with 400 uM indole (weak selection) or without Trp and with 25 uM indole (strong selection) for highly functional sequences (enrichment score > -5) (**D**) Distribution of mean enrichment scores (average of n=2 biological replicates) among the three selection conditions. (**E**) Hexbin plot of TransceptEVE score vs. measured mean enrichment score with strong selection. (**F**) Hexbin plot of TransceptEVE score vs. number of nonsynonymous substitutions for all sequences with a measured strong selection mean enrichment score.

concentration of indole (25 μM, strong selection). We tracked the abundance of library members in replicate yeast transformations of the same library over 4 timepoints taken at the beginning and end of 6 passages to obtain fitness scores. We found that fitness scores above a threshold (enrichment score > -5) were well correlated among replicates for both weak and strong selection conditions ($R^2$=0.58 and 0.65, respectively), but not for nonselective conditions where Trp was present (Fig. 3.5C), confirming the reliability of the assay for highly functional TrpB variants. Thousands of multi-mutation sequences at least as functional as the previously engineered high-fitness TrpB-003-1-A, with a $K_{cat}/K_M$ of $1.4 \times 10^5$ $M^{-1}s^{-1}$ [38], were identified (Fig. 3.5D). We also observed that a large fraction of sequences had low activity (Fig. 3.5D), which likely owes to the multi-copy nature of p1 that creates a delay in the action of purifying selection on recently generated mutants that hitchhike with functional TrpBs in the same cell.

We then asked whether a state-of-the-art ML model called TranceptEVE [69], which ensembles an autoregressive LLM (Tranception) trained across protein families with a variational autoencoder (EVE) trained on a specific family of proteins (in this case, TrpBs), could predict the measured fitness scores of our lab-evolved TrpBs. We found essentially no correlation between the relative fitness of TrpBs and their predicted fitness (Fig. 3.5E), even for the highly functional set of sequences (strong selection mean enrichment score >-5, Pearson correlation -0.064, p<0.0001), with the nearly nonfunctional *Tm*Triple variant being scored the highest among all sequences. In contrast and as expected, we found that scores exhibited a much stronger and negative correlation with the number of nonsynonymous amino acid mutations (Pearson correlation -0.433, p<0.0001, Fig. 3.5F). Our interpretation is that the evolution of TrpB for standalone function, rarely found in nature, combined with the extensiveness of sequence divergence from wt TrpB have brought our evolved sequences into regions of the fitness landscape that are out-of-distribution of natural sequences. It has been shown that large ML models can generate highly functional artificial sequences that are more dissimilar to natural sequences than our TrpBs are to wt TrpB [31]; it has also been shown that these models can nominate artificial sequences containing evolutionarily plausible mutations that improve non-natural functions [70]. Yet such ML-generated sequences are *ipso facto* within distribution of what was learned and may therefore miss important regions of underlying fitness landscapes. The ability for scalable continuous evolution to enter and explore functional regions of fitness landscapes that ML models do not, and vice-versa, highlights both open challenges in ML and the potential value of combining these two types of approaches going forward.

## 3.4 Discussion

In this work, we have engineered OrthoRep's mutation rate to reach $>10^{-4}$ s.p.b. while also reducing OrthoRep's bias against transversion mutations to maximize the exploration of sequence space. The durable action of these new mutation rates and preferences on a chosen gene, TrpB, enabled a new modality for continuous evolution where mutations quickly and durably accumulated on laboratory timescales even through periods of no selection or only purifying selection. In the absence of selection, TrpB sequences swiftly diffused into new regions of sequence space; under strong positive selection, new functional adaptations in TrpB rapidly emerged; and in the presence of purifying selection, TrpB sequences quickly sampled the space bounded by the constraints of structure and function, thereby revealing those constraints. At the end of our ~500 generation (<3 month) evolution experiment on TrpB, we obtained thousands of unique sequences, pairs of which were separated by an average ~35 amino acids (~9% divergence) including many >60 amino acids apart (>15% divergence). The amount of evolutionary information recorded into such extensive diversity allowed us to infer both known and unknown mechanisms shaping TrpB's evolution, including the focusing of adaptation on the COMM domain, the importance of certain allosterically linked positions on the function of TrpB, and the reduction of TrpB's isoelectric point to yield negatively charged variants even when selection for TrpB's enzymatic function was absent. The evolutionary information from the experiment also revealed structural and functional constraints acting on TrpB, including conservation of positions near the active site and in buried regions. Additionally, evolved outcomes included many highly active TrpBs that exceeded the *in vivo* fitness of previously evolved and engineered TrpBs. These variants were distinct from what could be predicted by ML models trained on natural proteins, indicating the discovery of high fitness regions in the fitness landscape of TrpB that are out-of-distribution of natural variation.

While TrpB was used to demonstrate the new capabilities of OrthoRep in both the evolutionary improvement of a gene's function and the extensive evolutionary recording of selective forces into sequence diversity, our experiments should easily extend beyond TrpB. The practicality of condensing long adaptive and neutral gene evolutionary processes into laboratory experiments should find broad applications in the evolutionary engineering of biomolecules, the mapping of sequence-function relationships, revealing novel biological constraints that shape evolution, and understanding how genes evolve — from their own points of view.

## 3.5 Materials and methods

### DNA plasmid construction

Plasmids used in this study are listed in Table S3.3, along with sources for DNA templates. Complete maps for these plasmids are made available for download on github at https://github.com/liusynevolab/OrthoRep_Rix_2023. All DNA templates for PCR were derived from previous studies or gBlocks (IDT). All primers were synthesized by IDT. All relevant primer pairs are listed in Table S3.4. Amplicons for construction of clonal plasmids were generated using Q5 Hot Start High-Fidelity DNA Polymerase (NEB). All non-library plasmids were constructed using Gibson Assembly and transformed into chemically competent *E. coli* strain TOP10 (ThermoFisher). Clonal plasmids were sequence verified by either Sanger sequencing (Azenta) or whole plasmid sequencing (Primordium).

### DNA library construction

Amplicons for TP-DNAP1 libraries were generated using error prone PCR with GeneMorph II (Agilent) according to manufacturer instructions, aiming for ~3-5 nucleotide substitutions per sequence. Amplicons for all other libraries were generated with Q5 Hot Start High-Fidelity DNA Polymerase (NEB). For epPCR 1, the resulting PCR product was assembled into plasmids using Gibson assembly in 20 µL reaction volumes. For all other libraries, resulting PCR products were assembled into plasmids with Golden Gate assembly with T4 DNA ligase and BsaI-HF v2 or PaqCI (all NEB) and in a 40 µL reaction volume. Gibson reactions were run at 50 °C for 1 hour. Golden gate reactions were run isothermally at 37 °C for 1 hour and heat inactivated at 65 °C for 10 minutes. Reactions were purified with AMPure XP beads (Beckman), typically with a 0.9:1 bead:sample ratio according to manufacturer instructions. Libraries were transformed into high-competency electrocompetent *E. coli* TOP10 cells (ThermoFisher).

### Yeast strains, media, transformations, and DNA extraction

All yeast strains used in this study and their provenance are listed in Table S3.5. Yeast were grown in liquid or on plates at 30 °C in synthetic complete (SC) growth medium (20 g/L dextrose, 6.7 g/L yeast nitrogen base w/ ammonium sulfate w/o amino acids (US Biological), appropriate nutrient drop-out mix (US Biological), as directed) or MSG SC growth medium (20 g/L dextrose, 1.72 g/L yeast nitrogen base w/o ammonium sulfate w/o amino acids (US Biological), appropriate nutrient drop-out mix (US Biological), as directed, 1 g/L L-Glutamic acid monosodium salt hydrate (ThermoFisher)) minus nutrients (referred to as -X where X is either the single letter amino acid code for an amino acid nutrient, or U for uracil) required for appropriate auxotrophy selection(s). Where selection for MET15 was required, cells were propagated in media lacking both methionine and cysteine. 500 µL liquid yeast cultures in 96-well deep well plates were incubated with shaking at 750 rpm. All other liquid yeast cultures were incubated with shaking at 200 rpm.

Yeast transformations, including p1 integrations and polymerase replacement integration, were performed as previously described[38]. For all integration transformations, plasmid DNA was linearized prior to transformation using either ScaI-HF or EcoRI-HF (both NEB) for p1 or genomic integrations, respectively. Due to its repetitive nature, deletion of FLO1 was performed by a URA3 knock-in knock-out method[71] (see Table S3.3, pFLO1-KO). Genetic deletions for TRP5 and MET15 were performed as previously described[72], using spacer sequences TTTGAGCCTGATCCCACTAG and GCTAAGAAGTATCTATCTAA, respectively. When isolating individual clones from genetic deletion and integration transformations, colonies were restreaked onto media agar plates of the same formulation to ensure isolation of only cells that have the desired genetic change.

All p1 plasmid sequences were generated by first generating a strain harboring a ‘landing pad’ p1 via integration and then integrating over this landing pad to generate the desired p1 construct. To enable construction of the landing pad strain, the wt TP-DNAP1 was integrated at the CAN1 locus using pGR475. A sequence encoding a partial LEU2 sequence lacking the N terminus was then integrated over the wt p1 using pGR420 to generate the landing pad p1. The wt p1, which encodes the TP-DNAP1, was then cured out via 3-4 1:1000 passages. The p1 plasmid(s) encoding the desired sequence were then generated via integration using cassette(s) that include a LEU2 sequence lacking the C terminus (*e.g.* pGR438). The overlap between the LEU2 on the landing pad and the new integration cassette was used to reconstitute full length LEU2 only when integration occured on p1, reducing the likelihood of genomic integration.

The polymerase replacement integration transformation was performed by first digesting 0.5-2 µg of the polymerase replacement plasmid or library with EcoRI-HF in a 25 µL reaction per 1x transformation followed by directly transforming this digestion reaction into a yeast strain encoding the CAN1-WT-TP-DNAP1 landing pad (all polymerase libraries were transformed into OR-Y488). Library scale transformations were carried out at 20-40x scale. Transformed yeast were plated onto solid MSG SC -LR or -MCR media w/ 100 mg/L nourseothricin (for positive selection of integration) and 200 mg/L L-canavanine (a toxic L-arginine analog for counterselection of cells that fail to perform polymerase replacement and remove the arginine permease CAN1). Leu or Met/Cys dropout was used to maintain selection for p1-encoded LEU2 or MET15, respectively, while Arg dropout was usedto improve L-canavanine selection.

Extraction of genomic DNA (gDNA) and p1/p2 plasmids was performed as previously described for 1.5 mL yeast culture volumes[38]. This procedure was used for all experiments except for DNA extracted for use in HTS dataset 6, which was instead performed in 96-well format for higher throughput. In brief, a 96-well block of 500 μL of saturated yeast cultures was centrifuged (2500 × G, 5 min), supernatant was discarded, pellets were resuspended in 1 mL 0.9% NaCl, this resupension was again centrifuged (2500 × G, 5 min), and the supernatant was discarded. The resulting pellet was resuspended in 250 μL Zymolyase solution (0.9 M D-Sorbitol (Sigma Aldrich), 0.1 M Ethylenediaminetetraacetic acid (EDTA, Sigma Aldrich), 10 U/mL Zymolyase (US Biological)) and incubated with shaking (37 °C, 200 RPM). The 96-well block was then centrifuged (2500 × G, 5 min), supernatant was discarded, and pellets were resuspended in 280.5 μL proteinase K solution (250 μL TE (50 mM Tris-HCl (pH 7.5), 20mM EDTA), 25 μL 10% sodium dodecyl sulfate (SDS, Sigma Aldrich), 5.5 μL proteinase K stock solution (10 mg/mL proteinase K (ThermoFisher)). The 96-well block was then incubated at 65 °C for 30 min, combined with 75 μL 5M potassium acetate (ThermoFisher), and incubated on ice for 30 min. The 96-well block was centrifuged at 12,000 × g for 10 min, the resulting supernatant was combined and mixed with 2 volumes buffer PB (5 M Guanidine hydrochloride (ThermoFisher), 30% isopropanol, 70% water), and this mixture was applied to a 96 well DNA-binding plate (Epoch Life Science) on a vacuum manifold. Flow through was discarded, columns were washed with PE buffer (10 mM Tris-HCl (ThermoFisher), 80% ethanol, 20% water, pH 7.5), centrifuged and dried, and 60 μL water was applied to columns for elution by centrifugation (2500 × G, 5 min).

### High throughput sequencing

All high throughput sequencing datasets are listed in Table S3.6, along with the method used to construct them. All PCRs for high throughput sequencing were performed with Platinum SuperFi II DNA Polymerase (ThermoFisher). For short read paired end sequencing, both low cost/yield (AmpliconEZ, Azenta) and high cost/yield (HiSeq paired end 150, Novogene) were performed directly on PCR products generated in either one or two rounds of PCR using primers that each included an adapter sequence, a 6- or 7-nucleotide barcode, or both.

For in-house long read sequencing, we used the Oxford Nanopore Technologies nanopore sequencing platform. Due to the lower accuracy of nanopore sequencing, we employed modified versions of previously described methods for the construction of DNA libraries that yield multiple reads of the same original DNA molecule, allowing for computational reconstruction of high accuracy sequences[54,73]. We refer to the first of these as *in vivo* downsampled unique molecular identifier (UMI) PCR, which was used for most nanopore sequencing in this study (Table S3.6). This involved first a 2-cycle “UMI tagging” reaction, in which primers (*e.g.*, primer pair 2) were used to append both UMIs and universal DNA sequences (for further amplification) to both ends of the target sequence with the following components:

- ~1-50 ng of purified yeast or *E. coli* miniprep
- 5 µL 2x SuperFi II master mix
- 1 µM each primer
- water to 10 µL

and with these thermocycler conditions:

1. 98 °C for 30 sec
2. 98 °C for 10 sec
3. 65 °C for 1 sec
4. 60 °C for 45 sec w/ ramp down from 65 to 60 @ 0.2 °C per second (this should result in 25 seconds of ramp down time and 20 seconds of hold time)
5. 72 °C extension, 1 min / kb

6. Go to step 2 (1x)
7. 72 °C for 2 min

Next, UMI primers were removed using ExoSAP-IT (ThermoFisher) according to manufacturer instructions. 10 µL of the resulting reaction was then used as template in a 25 µL Platinum SuperFi II PCR with 1 mM $MgCl_2$ supplemented, using primers (*e.g.*, primer pair 3) that included BsaI or PaqCI sites and 7 nucleotide barcodes in forward/reverse combinations that were unique for each sample. The resulting uniquely barcoded PCR products were then combined, purified using AMPure XP beads, used in a library Golden Gate reaction with an *E. coli* vector, then transformed into high competency *E. coli* as described above. Plasmid pGR554 (Table S3.3) was designed for this purpose and contains both CcdB and sfGFP, which both are replaced with the insert during Golden Gate assembly, as well as NotI and SbfI sites, strategically placed to enable separation of the desired library insert from the backbone prior to sequencing. CcdB and sfGFP provided counterselection and visualization of colonies resulting from undigested vector. (For reference, cloning into any *E. coli* vector suffices, so long as unique library members are associated with a unique relatively short (20-50 bp) sequence.) Resulting colonies each contained many copies of a unique plasmid species encoding a UMI-tagged library member. To obtain good coverage of each sequence and UMI with nanopore sequencing, the resulting library was downsampled by only harvesting ~20-fold fewer colonies than the expected number of reads, amounting to 100 – 200 thousand colonies for a standard MinION flow cell. Plasmid DNA from this library was then miniprepped, digested (for example with NotI-HF or NotI-HF/SbfI-HF (both NEB) if using plasmid pGR554), and gel extracted prior to sequencing.

Construction of the library for HTS dataset 8a was performed using a similar approach, but with a yeast expression vector, and with UMIs and library members inserted into the vector in two distinct steps so that UMIs are present at a single location in the plasmid and would not need to be immediately adjacent to library members, mitigating any potential effects of UMIs on expression. In brief, libraries of UMIs generated via PCR using primer pair 8 were cloned into plasmid pUMI by Golden Gate assembly with BsmBI-v2 and *E. coli* electrotransformation to generate an intermediate UMI library. Different variants of pUMIs with unique, known 7-nucleotide barcodes were used for each library to allow multiplexing. These libraries were then used as the vector into which evolved TrpB library members, PCR amplified using primer pair 9, were cloned via standard restriction cloning, with inserts digested with PaqCI and vectors digested with BsaI-HFv2. Isothermal ligation (1 hour 37 °C, 20 mins 65 °C) was then performed with T4 ligase in 40 µL reaction volumes, AMPure bead purified, and transformed into high efficiency *E. coli*. This was carried out for 6 unique libraries: downsampled in yeast with selection, downsampled in yeast without selection, and not downsampled in yeast, each for both timepoints (generation 350 and generation 540) chosen from TrpB evolution. The intermediate UMI libraries were generated at a library size of >100-fold larger than the desired final library size to minimize the chance that distinct library members would be tagged with identical UMIs.

The second method used for generating libraries for high accuracy nanopore sequencing was adapted from methods described in Volden *et al.*[73], Oliynyk and Church[74], and Zhang and Tanner[75]. It involved circularization of the target sequence and use of this circularized product as template for rolling circle amplification (RCA) using strand displacing DNA polymerases (Fig S15). In brief, UMI-tagged PCR products were generated as described above, albeit with complementary Type IIS cut sites (BsaI or PaqCI) on the forward and reverse primers used during PCR amplification such that the two ends of the amplicon ligated to each other during a Golden Gate assembly reaction, forming a circular product. Due to the large amount of template DNA required for rolling circle amplification, a large amount of amplicon was used in the circularization reaction, typically 1-2 pmols in 100 µL Golden Gate assembly reactions. Oliynyk and Church[74] provide a useful discussion of relevant considerations for such circularization reactions. An isothermal Golden Gate assembly reaction was then performed to circularize the amplicon library.

Following Golden Gate assembly, uncircularized DNA was digested with lambda exonuclease (NEB), exonuclease I (NEB), and exonuclease III (NEB) in a 1:1:0.1 ratio, which was added directly to the Golden Gate reaction in a 1:10 exonuclease mixture to sample ratio. This reaction was incubated at 37 °C for 45 min, then 80 °C for 15 min. The reaction was then AMPure bead purified with a 0.7:1 bead:sample ratio, eluting in a maximum of 10 µL of water, and the resulting circularized library was used in a rolling circle amplification reaction with the following components combined on ice:

- 4 µL 10X NEB buffer 4 (NEB)
- 4.8 µL 10 mM dNTPs
- 2.64 µL 5 U/µL Bsu DNA Polymerase, Large Fragment (NEB)
- 1.6 µL 10 µg/µLT4 gene 32 protein (NEB)
- 0.2 – 1 µg circularized DNA library
- RCA primers, 2 µM each (must bind internal to first primer set, *e.g.* primer pair 7)
- Water to 40 µL

This reaction was incubated at 37 °C for 3 hours. SDS-containing loading dye was added directly to the reaction, and the entire sample was run on an agarose gel. Bands corresponding to 3x–6x concatemers were gel extracted, and purified DNA was used for nanopore sequencing. Unlike RCA using Phi29, this method enabled size selection of specific repeat numbers and did not require a ‘debranching’ step, but required large amounts of input DNA and in our experience suffered from high sensitivity to DNA contamination. Use of Phi29 with random hexamers, followed by debranching, is therefore a reasonable alternative.

Following construction of *in vivo* downsampled UMI PCR or RCA libraries, nanopore library preparation and sequencing was performed using the most up-to-date ligation sequencing kit (*e.g.* LSK-114) and flow cell (*e.g.* R10.4.1), following manufacturer instructions, with two exceptions. First, ½ volumes (but unaltered DNA input) for end prep and ligation reactions were used and second, FFPE Repair Mix was not used during end prep reactions.

All relevant high throughput sequencing datasets will be made available on the NCBI sequence read archive (SRA) prior to peer-reviewed publication.

## Error rate measurement by mutation accumulation

Following a polymerase replacement integration transformation, colonies were picked into liquid media of the same media formulation as that used for selection after transformation, and then grown to saturation. The resulting saturated culture was miniprepped to serve as the 0th passage, $p_0$. The culture was then also propagated in the appropriate growth medium for maintenance of the orthogonal plasmid, typically SC -L, using a dilution factor $d$ (typically 128, 256, or 512) that is consistent throughout the experiment. At least one additional saturated culture from this time course was miniprepped to serve as the $t^{th}$ passage, $p_t$. We used only two passages/timepoints in all mutation accumulation error rate measurement experiments except that which produced HTS dataset 6. The number of generations, $g$, that separated $p_0$ from $p_t$ was used to calculate the mutation rate and can be approximated as

$$g_t = t \times \log_2 d\,.$$

We note that this approximation assumes no cell death and equivalent saturation at each passage. Cultures were manually propagated several times until a total number of generations of at least 50 was reached. Miniprepped yeast DNA for each timepoint was used as template DNA for high throughput sequencing, and custom scripts, organized within the Maple pipeline, were used to calculate mutation rates.

Mutation rates were calculated as the rate of accumulation of a mutation type (all substitutions, individual substitutions, insertions, or deletions), *j*. We denote $\mu_j$ as this rate calculated for mutation type *j*. High accuracy HTS was used to first obtain $c_{j,t}$, the total counts of mutation *j* (*e.g.*, A to T) among all sequences in passage *t*. For substitution mutations, to account for the influence of variable A/T/G/C content in the sequence being analyzed, this count is normalized to obtain $n_{j,t}$, the expected count for an idealized sequence with a 1:1:1:1 A:T:G:C ratio. For a substitution at a particular nucleotide where the nucleotide occurs *w* times within a reference sequence of length *L*, $n_{j,t}$ is calculated as

$$n_{j,t} = \frac{c_{j,t}}{w} \times \frac{L}{4}\,.$$

The total normalized count of all substitution mutations at each timepoint is then calculated as the sum of all $n_{j,t}$ for all twelve substitution types. However, for insertions and deletions, $n_{j,t}$ is not normalized in this way, and is instead equivalent to $c_{j,t}$, the total number of nucleotides inserted or deleted among all sequences for that timepoint.

To obtain the per-nucleotide, per-generation mutation rate $\mu_j$, we used the total number of sequences analyzed for each timepoint, $s_t$, to calculate the per-nucleotide frequency of mutation *j* for each timepoint. We then use linear regression on these normalized per-nucleotide frequencies according to

$$\frac{n_{j,t}}{L \times s_t} = \mu_j \times g_t + b_j,$$

where $\mu_j$ and $b_j$ are the slope and intercept, respectively, of the best fit line for all *t* timepoints. When the number of generations between the 0th timepoint and the initiation of mutagenesis (typically polymerase replacement) is accurately estimated and no mutations fully fixed within the population prior to the first timepoint, $b_j$ should be close to 0. Regardless, we do not report $b_j$, as it has no bearing on $\mu_j$ across experiments. We report $\mu_j$ as the per-base per-generation rate of accumulation of mutation type *j*. Mutation tabulation was performed by the script mutation_analysis.py and all other operations related to mutation rate calculation were performed by the script plot_mutation_rates.py, both of which are contained within the Maple pipeline. All reported mutation rates were calculated using mutation tabulation within a sequence region that was not under functional selection.

## TP-DNAP1 library selection and screening

Following TP-DNAP1 replacement library construction, resulting purified plasmid DNA was used for a polymerase replacement integration transformation. Prior to transformation, OR-Y488 was grown up in SC -L + 1 mg/L 5-fluoroorotic acid (US Biological) for counterselection against cells that had reverted the inactivating mutation in ura3* by chance. Following library transformation,plating on media selecting for cells that had replaced the wild type TP-DNAP1 with library variants, and 48 hour incubation, colonies were harvested in bulk and immediately plated onto SC -LU plates. These plates were then incubated for 48 hours, and resulting colonies were either picked into liquid media for mutation rate or frequency characterization (by mutation accumulation or fluctuation assays, respectively) or were harvested in bulk and immediately plated onto SC -LUW media. After colony formation, colonies were either picked into liquid media for mutation rate characterization by mutation accumulation or were harvested in bulk, miniprepped for gDNA isolation, and used as template for a non-mutagenic PCR to generate amplicons for Golden Gate assembly into pGR554, which was then retransformed into OR-Y488 to repeat the selection and perform mutation rate screening.

The fluctuation test was performed as follows. Following transformation of the epPCR 1 TP-DNAP1 library into OR-Y488 and selection for ura3* reversion on solid media, individual colonies were picked from this plate, inoculated into 500 µL SC -LU media in a 96 well block, and grown to saturation. Cultures were then passaged 1:10,000 into 200 µL SC -LU, 12 replicates per each individual colony, and grown to saturation. Cultures were centrifuged, washed with 0.9% NaCl, and pellets were resuspended in 35 µL 0.9% NaCl. Each replicate 10 µL of this resuspension was plated onto SC -LUW plates. A subset of cultures were titered and plated on SC -LU plates to estimate population size. Plated cells were allowed to grow for 4 days, and revertants on each spot were counted. Counts were used to estimate the *m* value using the FALCOR online web tool (https://lianglab.brocku.ca/FALCOR/) and the Ma-Sandri-Sarkar Maximum Likelihood Estimator. Mutation frequency was calculated from this *m* value as previously described[42], using a target size of 1 (only one mutation is capable of restoring Trp5 activity). Copy number was not considered and therefore per base substitution rate was not calculated using this method.

## TrpB evolution

Plasmid pGR595 (TP-DNAP1 BadBoy2) was first transformed into yeast strain OR-Y484 according to the polymerase replacement procedure described above. The resulting strain (OR-Y538) was then transformed with plasmid pGR438 (TrpB with lineage barcodes), following the p1 integration procedure described above, plating on SC -L media. All ~400 resulting colonies were harvested together and passaged into 512 µL SC -L media in all wells of a 96-well block and grown to saturation. DNA extracted from these resulting cultures served as passage/timepoint 0, which we approximate to be ~50 generations from TrpB p1 integration. These cultures were also passaged 1:1024 (0.5 µL into 512 µL) for all passages in the experiment into growth media and with timepoints taken as described in Fig. S14. DNA extraction for timepoints was performed by combining all 96 saturated cultures for a specific timepoint and extracting DNA from the pooled cultures according to the 1.5 mL DNA extraction protocol.

To downsample evolved populations for cloning UMI-tagged TrpB libraries, yeast cultures from the passages corresponding to generation 340 and 510 were inoculated into SC -L media from glycerol stock and grown to saturation. Saturated cultures were then combined, and multiple serial dilutions of both cultures were plated onto SC -L media. Plates derived from generation 340 and 510 with ~3700 and ~1700 colonies respectively were harvested. These cultures were passaged 1:1000 into 2 mL of either SC -LW + 400 µM indole (Sigma-Aldrich) (selective) or SC -L (nonselective) growth media and allowed to grow to saturation. This process was repeated twice more for each of the two media. All three of the resulting saturated cultures for each selective and nonselective conditions were pooled and DNA was extracted to serve as template for cloning the TrpB library fitness assay library, in addition to DNA extracted from generation

340 and 510 without downsampling. Note that this downsampling performed in yeast preceded the downsampling in *E. coli* necessary for proper sequencing coverage.

### Pooled TrpB fitness assay

UMI-tagged evolved TrpB libraries and equivalent plasmids expressing two control TrpB sequences (*Tm*Triple and TrpB-003-1-A) were transformed into yeast strain OR-Y260 and plated onto -LH media, resulting in 40-fold coverage of the ~120 thousand member library. Colonies were harvested and the library was spiked with each of the two control TrpB-expressing yeast strains at a 1:1000 control:library ratio. DNA was extracted from the resulting library to serve as the 0th timepoint. This library was also passaged 1:100 into 50 mL of either SC -L (nonselective), SC -LW + 400 μM indole (weakly selective), or SC -LW + 25 μM indole (strongly selective) and grown to saturation. This passaging and growth was repeated for each of the three growth conditions five times for a total of six passages. Of these, DNA was extracted from passages 1, 2, 5, and 6 to serve as additional timepoints. DNA from all timepoints were used as templates for PCR amplification of only the UMI and barcode regions for high throughput sequencing.

Enrichment scores were calculated following the procedure described in Rubin *et al.*[76], albeit with two modifications to account for disparate sequencing coverage over multiple timepoints. First, counts of each UMI at each timepoint were normalized to the average count of all UMIs that persisted throughout all timepoints prior to log transformation and weighted linear regression. Second, regression weights were multiplied by this average count. All enrichment calculations were performed by the Maple pipeline within the script enrichment.py.

## High-throughput sequencing analysis

High accuracy consensus sequence generation, alignment, demultiplexing, genotype identification, mutation rate analysis, and basic dataset visualization were performed by version v0.10.4 of the Maple pipeline, which uses the Snakemake workflow management library, and is available on https://github.com/gordonrix/maple. Parameters and settings for Maple analyses for each dataset and all other code used for analysis are available on Github (https://github.com/liusynevolab/OrthoRep_Rix_2023).

## TrpB fitness prediction with TranceptEVE[69]

A multiple sequence alignment (MSA) of natural TrpB subunits was created using 5 iterations of jackhmmer[77] to query the UniRef100 database with a bitscore of 0.9. Columns with more than 20% gaps were ignored, and we used a theta parameter of 0.8 to downweight sequences with more than 80% sequence homology, as described in Hopf *et al.*[78]. We trained 4 separate EVE models using this MSA and used them to calculate the log probabilities of the mutated sequences. These scores were ensembled with the Tranception log probabilities for each sequence by taking a weighted average of 60% Tranception score and 40% EVE score.

## Null hypothesis sequence simulation and analysis

A dataset of TrpB variants that contained random mutations representative of biases due to the relevant mutation preferences and wild type sequence, but that were not subject to selective forces, was generated through a simple simulation. First, the number of synonymous mutations within the ORF of each unique evolved TrpB sequence was approximated as the number of nucleotide mutations minus the number of nonsynonymous mutations. For each real unique genotype, a corresponding simulated genotype was generated by starting from the wild type TrpB sequence used in the evolution experiment and stochastically sampling nucleotide mutations with probabilities determined by the mutation rates of the same polymerase used for TrpB evolution (BadBoy2) until the same number of synonymous mutations as the evolved genotype was reached. Additional information such as the timepoint from which the real sequence was identified and the count of the genotype was also replicated for the corresponding simulated sequence. To account for some minor strand-dependent mutational biases, mutation rates and spectrum calculated for a sequence in the same position and orientation (relative to the LEU2 gene) as TrpB were used to generate the simulated sequences.

Simulated and evolved sequences were analyzed identically. Isoelectric point and hydrophobicity index (gravy) were both calculated using the ProtParam module in BioPython[79]. Mesophilic adaptation mutations were selected from Haney *et al.*[13] as the inverse of the 17 mesophile to thermophile “Replacements most biased in number” with P<0.005. Alternative sets of amino acid replacements were not evaluated.

## Computational lineage downsampling

Lineage barcodes were identified using the demultiplexing feature within the Maple pipeline. The 100 most frequently observed lineage barcodes only from the first timepoint of TrpB evolution were used for all lineage analyses. Lineage barcodes for all remaining timepoints were assigned from this list of 100 barcodes, allowing for a nucleotide hamming distance of 1. For global covariation analysis, all barcodes appearing in at least 100 sequences were identified, and 100 sequences from each lineage were randomly extracted for analysis of mutation frequencies. Covariation with specific mutations was performed similarly, except that only sequences that contained the specific mutation and were identified from specified timepoints were considered prior to randomly extracting sequences for analysis of mutation frequencies. To ensure random sampling did not bias results, this process was performed multiple times with virtually identical results.

## 3.6 References

1. Blow, D. M., Birktoft, J. J. & Hartley, B. S. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature.* **221**, 337–340 (1969).

2. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178 (1995).

3. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

4. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **39**, 37–43 (2011).

5. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell.* **138**, 774–786 (2009).

6. McCormick, J. W., Russo, M. A. X., Thompson, S., Blevins, A. & Reynolds, K. A. Structurally distributed surface sites tune allosteric regulation. *eLife.* **10**, 1–38 (2021).

7. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, (2011).

8. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, (2011).

9. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* **286**, 295–299 (1999).

10. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods.* **14**, 45–48 (2016).

11. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in

protein structures be predicted by analysis of correlated mutations? *Protein Eng. Des. Sel.* **7**, 349–358 (1994).

12. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707 (1987).

13. Haney, P. J. *et al.* Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3578–3583 (1999).

14. Gianese, G., Argos, P. & Pascarella, S. Structural adaptation of enzymes to low temperatures. *Protein Eng.* **14**, 141–148 (2001).

15. Berezovsky, I. N. & Shakhnovich, E. I. Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12742–12747 (2005).

16. Marcotte, E. M., Xenarios, I., Van Der Bliek, A. M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 12115–12120 (2000).

17. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. Conserved residues and the mechanism of protein folding. *Nature.* **379**, 96–98 (1996).

18. Wolfenden, R. V., Cullis, P. M. & Southgate, C. C. F. Water, Protein Folding, and the Genetic Code. *Science.* **206**, 575–577 (1979).

19. Collias, D. & Beisel, C. L. CRISPR technologies and the search for the PAM-free nuclease. *Nat. Commun.* **12**, 1–12 (2021).

20. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A Panel of TrpB Biocatalysts Derived from Tryptophan Synthase through the Transfer of

Mutations that Mimic Allosteric Activation. *Angew. Chemie - Int. Ed.* **55**, 11577–11581 (2016).

21. Baier, F. *et al.* Cryptic genetic variation shapes the adaptive evolutionary potential of enzymes. *eLife.* **8**, 1–20 (2019).

22. Gasiunas, G. *et al.* A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat. Commun.* **11**, (2020).

23. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).

24. Trudeau, D. L., Smith, M. A. & Arnold, F. H. Innovation by homologous recombination. *Curr. Opin. Chem. Biol.* **17**, 902–909 (2013).

25. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature.* **391**, 288–291 (1998).

26. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods.* **15**, 816–822 (2018).

27. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature.* **599**, 91–95 (2021).

28. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

29. Shin, J. E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 1–11 (2021).

30. Bryant, D. *et al.* Massively parallel deep diversification of AAV capsid proteins by

machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).

31. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-022-01618-2.

32. Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell.* **149**, 1607–1621 (2012).

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

34. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature.* **596**, 590–596 (2021).

35. Makałowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 9407–9412 (1998).

36. Nei, M., Xu, P. & Glazko, G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2497–2502 (2001).

37. Stiffler, M. A. *et al.* Protein Structure from Experimental Evolution. *Cell Syst.* **10**, 15-24.e5 (2020).

38. Rix, G. *et al.* Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nat. Commun.* **11**, 1–11 (2020).

39. Morrison, M. S., Podracky, C. J. & Liu, D. R. The developing toolkit of continuous directed evolution. *Nat. Chem. Biol.* **16**, 610–619 (2020).

40. Molina, R. S. *et al.* In vivo hypermutation and continuous evolution. *Nat. Rev. Methods*

*Prim.* **2**, (2022).

41. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

42. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell.* **175**, 1946-1957.e13 (2018).

43. Zhong, Z., Ravikumar, A. & Liu, C. C. Tunable Expression Systems for Orthogonal DNA Replication. *ACS Synth. Biol.* **7**, 2930–2934 (2018).

44. García-García, J. D. *et al.* Using continuous directed evolution to improve enzymes for plant applications. *Plant Physiol.* (2021) doi:10.1093/plphys/kiab500.

45. Jensen, E. D. *et al.* Integrating continuous hypermutation with high-throughput screening for optimization of cis,cis-muconic acid production in yeast. *Microb. Biotechnol.* **14**, 2617–2626 (2021).

46. Javanpour, A. A. & Liu, C. C. Evolving Small-Molecule Biosensors with Improved Performance and Reprogrammed Ligand Preference Using OrthoRep. *ACS Synth. Biol.* **10**, 2705–2714 (2021).

47. Wellner, A. *et al.* Rapid generation of potent antibodies by autonomous hypermutation in yeast. *Nat. Chem. Biol.* **17**, 1057–1064 (2021).

48. Harvey, E. P. *et al.* An in silico method to assess antibody fragment polyreactivity. *Nat. Commun.* **13**, (2022).

49. Vallina Estrada, E., Zhang, N., Wennerström, H., Danielsson, J. & Oliveberg, M. Diffusive intracellular interactions: On the role of protein net charge and functional adaptation. *Curr. Opin. Struct. Biol.* **81**, (2023).

50. Brown, T., Hunter, W. N., Kneale, G. & Kennard, O. Molecular structure of the G·A base pair in DNA and its implications for the mechanism of transversion mutations. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 2402–2406 (1986).

51. Luria, S. E. & Delbrück, M. Mutations of Bacteria From Virus Sensitivity To Virus Resistance. *Genetics* **28**, 491–511 (1943).

52. Foster, P. L. Methods for Determining Spontaneous Mutation Rates. *Methods Enzymol.* **409**, 195–213 (2006).

53. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 1–11 (2016).

54. Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nat. Commun.* **11**, 1–10 (2020).

55. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods.* **18**, 165–169 (2021).

56. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* **178**, 67–82 (2008).

57. Tian, R. *et al.* Engineered bacterial orthogonal DNA replication system for continuous evolution. *Nat. Chem. Biol.* (2023) doi:10.1038/s41589-023-01387-2.

58. Orr, H. A. The Rate of Adaptation in Asexuals. **968**, 961–968 (2000).

59. Gerrish, P. J., Colato, A. & Sniegowski, P. D. Genomic mutation rates that neutralize adaptive evolution and natural selection. *J. R. Soc. Interface* **10**, (2013).

60. Dunn, M. F. Allosteric regulation of substrate channeling and catalysis in the tryptophan

synthase bienzyme complex. *Arch. Biochem. Biophys.* **519**, 154–166 (2012).

61. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci.* **112**, 14599–14604 (2015).

62. Buller, A. R. *et al.* Directed Evolution Mimics Allosteric Activation by Stepwise Tuning of the Conformational Ensemble. *J. Am. Chem. Soc.* **140**, 7256–7266 (2018).

63. Maria-Solano, M. A., Iglesias-Fernández, J. & Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **141**, 13049–13056 (2019).

64. Orij, R., Postmus, J., Beek, A. Ter, Brul, S. & Smits, G. J. In vivo measurement of cytosolic and mitochondrial pH using a pH-sensitive GFP derivative in Saccharomyces cerevisiae reveals a relation between intracellular pH and growth. *Microbiology* **155**, 268–278 (2009).

65. Wennerström, H., Estrada, E. V., Danielsson, J. & Oliveberg, M. Colloidal stability of the living cell. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10113–10121 (2020).

66. Xiang, L., Yan, R., Chen, K., Li, W. & Xu, K. Single-Molecule Displacement Mapping Unveils Sign-Asymmetric Protein Charge Effects on Intraorganellar Diffusion. *Nano Lett.* **23**, 1711–1716 (2023).

67. Leander, M., Yuan, Y., Meger, A., Cui, Q. & Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25445–25454 (2020).

68. Shaw, A., Spinner, H., Shin, J., Gurev, S. & Rollins, N. Removing bias in sequence models of protein fitness. (2023).

69. Notin, P. *et al.* TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. *bioRxiv* 2022.12.07.519495 (2022).

70. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01763-2.

71. Alani, E., Cao, L. & Kleckner, N. A method for gene disruption that allows repeated use of URA3 selection in the construction of multiply disrupted yeast strains. *Genetics* **116**, 541–545 (1987).

72. Ryan, O. W. & Cate, J. H. D. Multiplex engineering of industrial yeast genomes using CRISPRm. in *Methods in Enzymology* vol. 546 473–489 (Elsevier Inc., 2014).

73. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9726–9731 (2018).

74. Oliynyk, R. T. & Church, G. M. Efficient modification and preparation of circular DNA for expression in cell culture. *Commun. Biol.* **5**, 1–10 (2022).

75. Zhang, Y. & Tanner, N. A. Isothermal Amplification of Long, Discrete DNA Fragments Facilitated by Single-Stranded Binding Protein. *Sci. Rep.* **7**, 1–9 (2017).

76. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 1–15 (2017).

77. Johnson, L. S., Eddy, S. R. & Portugaly, E. Johnson, L.S.; Eddy, S.R.; Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinform. 2010, 11, 431. *BMC Bioinformatics* **11**, 431 (2010).

78. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).

79. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

## 3.7 Author list

Gordon Rix, Rory L. Williams, Hansen Spinner, Vincent J. Hu, Debora S. Marks, and Chang C. Liu

## 3.8 Author contributions

Conceptualization: GR, CCL

Methodology: GR, CCL, VJH, HS

Investigation: GR, RLW, HS

Visualization: GR

Funding acquisition: CCL, DSM

Supervision: CCL, DSM

Writing – original draft: GR, CCL

Writing – review & editing: GR, CCL, RLW, HS, DSM

# Chapter 4. Developing a direct selection for *in vivo* noncanonical amino acid production

## 4.1 Introduction

A remarkable diversity of protein functions are accomplished using the 20 standard proteinogenic amino acids, which comprise only a fraction of the virtually limitless chemical diversity of noncanonical amino acids (ncAAs). Unsurprisingly, this chemical diversity can serve to expand the scope of protein function. Site-specific installation of ncAAs that serve as handles for bioorthogonal chemistries[1], photocaged reactive groups for light-induced catalysis[2], and replacing solublized catalysts[3]. These applications take advantage of orthogonal translation systems, in which a tRNA synthetase (tRS) has been engineered to use a ncAA to charge its cognate tRNA with its anticodon typically altered to recognize the amber stop codon. Critically, such tRNA / tRS pairs operate orthogonally to the host pairs such that the tRNA is not recognized by native tRSs and the tRS does not recognize native tRNAs, enabling the site specific incorporation of only available ncAAs that are recognized by the tRS.

Unfortunately, site-specific ncAA incorporation as well as the myriad other applications of ncAAs are stymied by difficulties in their synthesis, owed largely to the typically chiral α-carbon, resulting in their limited availability and typically high cost. Enzymes have no qualms with carrying out precise stereoselective chemistry, and have therefore proven their worth in mitigating these difficulties.[4] Beyond their application in *ex vivo* ncAA synthesis, enzymes that enable *in vivo* synthesis of ncAAs from readily available chemical precursors may facilitate biologic drug synthesis and site-specific ncAA incorporation into proteins.[5]

Enzymes are rarely built for the task of synthesizing unnatural amino acids in their natural form. The use of directed evolution for improving such activities has therefore proven very valuable for engineering enzymes for this purpose[4,6]. However, directed evolution campaigns for improving enzymatic ncAA synthesis typically depend upon directly screening individual enzyme variants for the desired activity. This provides exquisite control over reaction conditions and precise, high resolution measurements of enzyme activity, but comes at the cost of throughput. Selections and high throughput screens instead employ a linkage between cell survival or a single-cell-resolution optical signal and the desired activity to enrich for desired enzyme variants in high throughput. The implementation of selections and high throughput screens therefore have the potential to improve the speed at which enzymes can be engineered for ncAA production.

In this work, we establish a system for the growth-based selection of enzymes with increased ncAA production using an orthogonal tRNA / tRS pair to suppress stop codons in a selectable marker. This selection is versatile. Any enzymes that produce ncAAs recognized by tRNA / tRS pairs that are orthogonal in yeast are potential targets. We apply this selection to the *Thermatoga maritima* tryptophan synthase β-subunit, *Tm*TrpB, resulting in the isolation of enzymes capable of producing detectable amounts of 3-iodo-L-tyrosine *in vivo* at. Further development of this selection platform, and its pairing with *in vivo* hypermutation systems, should enable the *in vivo* production of diverse ncAAs.

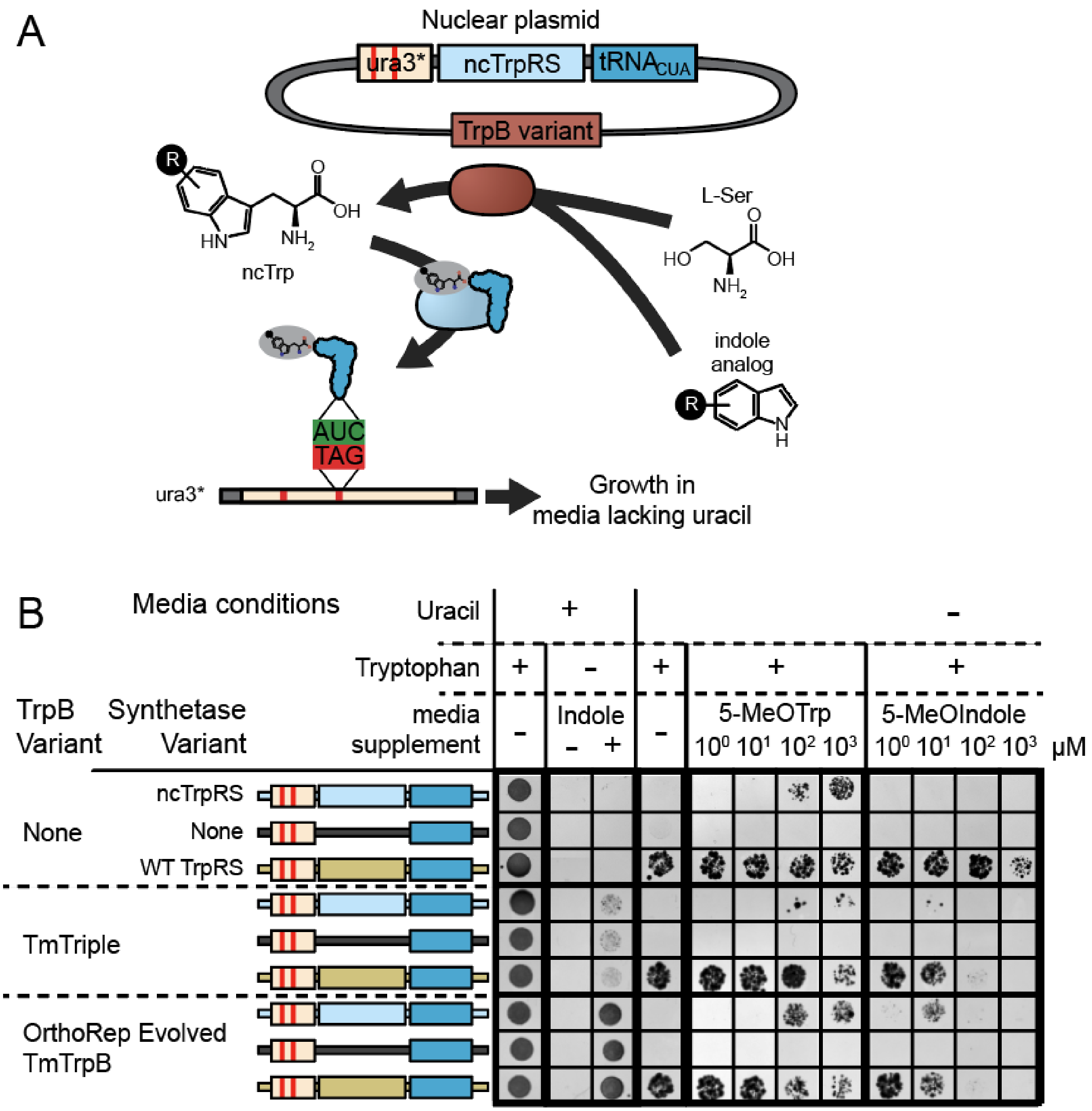


**Figure 1. A growth-based selection for noncanonical Trp analogs (ncTrp) via an expanded genetic code.** A) Illustration of the genetic circuit. A TrpB variant that is capable of producing a noncanonical Trp analog (ncTrp) enables charging of an amber codon suppressor tRNA (tRNA$_{CUA}$) by an engineered tryptophanyl tRNA synthetase (ncTrp tRS). The resulting charged tRNA enables suppression of two stop codons that are installed into the conditionally essential ura3 gene. Growth in media lacking uracil is therefore rendered dependent upon the production of ncTrp by the cell. B) Validation of the growth-based ncTrp analog selection. Images of yeast encoding the indicated genetic construct (left) plated onto the indicated growth medium (top) and grown for 3 days.

## 4.2 Results and discussion

### A growth-based selection for alternative Trp activities

*Tm*TrpB natively produces L-tryptophan (Trp) in a condensation reaction between indole and L-serine. We previously used OrthoRep, an *in vivo* hypermutation system in *S. cerevisiae*, to engineer variants of *Tm*TrpB that function well in yeast and were capable of producing noncanonical Trp analogs (ncTrp), such as 5-MethoxyTrp (5MeOTrp), *in vitro* (see Chapter 1). We therefore reasoned that this activity could be used to establish and test a selection for *in vivo* ncAA production. The native *E. coli* tryptophanyl tRNA / tRS pair was previously engineered for orthogonal translation in mammalian cells, yielding a tRNA synthetase capable of charging an amber codon suppressor variant of its cognate tRNA with a handful of 5-substituted ncTrp species, among them 5MeOTrp[7]. Reasoning that this synthetase would likely also exhibit orthogonality in yeast, we chose this tRS, hereafter ncTrp tRS, to validate our selection.

We designed a selection circuit using this tRNA / tRS pair for *in vivo* ncTrp production via amber stop codon suppression in URA3, an enzyme essential for uracil biosynthesis (Figure 1A). We expressed ncTrp tRS under a high strength constitutive promoter, and expressed its cognate stop suppressor tRNA  using a dicistronic tRNA operon that was previously reported to confer high expression of stop suppressor tRNAs.[8] An expression cassette for the native *E. coli* TrpRS, WT TrpRS, was also included as a positive control. WT TrpRS can accept Trp and should therefore confer uracil prototrophy in the absense of ncTrp. We also expressed URA3, into which we had installed two TAG amber stop codons at sites that are poorly conserved among URA3 orthologs. Finally, we expressed a *Tm*TrpB variant that had been previously engineered to exhibit standalone activity *in vitro* when expressed in *E. coli* (*Tm*Triple[9]) or *in vivo* in yeast (Q90*-003-1-A, Chapter 1). We tested this selection circuit via plating assays on various growth media (Figure 1B). Cells expressing either *Tm*Triple or WT-100-A were capable of surviving in media lacking uracil only in the presence of 5-MeO-indole, demonstrating direct selection for *in vivo* Trp analog production. Furthermore, the concentration of 5-MeO-indole required for uracil prototrophy was 10-fold lower than that of 5-MeOTrp, suggesting that higher intracellular 5-MeOTrp concentrations can be achieved through *in vivo* biosynthesis, perhaps due to better cell permeability of 5-MeOTrp. The TrpB-dependent elevated cytotoxicity of 5-MeOIndole provides further evidence for this hypothesis. *In vivo* biosynthesis of 5-MeOTrp, and likely other Trp analogs, may therefore offer an improvement over exogenous supplementation of the ncAA beyond the reduced cost.

These efforts demonstrated that growth-based selections that directly enrich for production of ncAAs by TrpB may have utility in further engineering such activities. However, the ncAAs that this engineered ncTrp tRS could utilize efficiently, such as 5-MeOTrp, are already produced *in vitro* at high yields by *Tm*TrpB variants we had previously evolved (see Chapter 1). Not only does this reduce the industrial value of engineering TrpBs with improvements to their production of such ncAAs, but, given OrthoRep's ability to traverse long evolutionary trajectories, more dramatic evolutionary transformations are likely achievable through selections for weaker TrpB activities. We therefore sought to develop a similar approach that would enable selection for production of noncanonical tyrosine analogs.

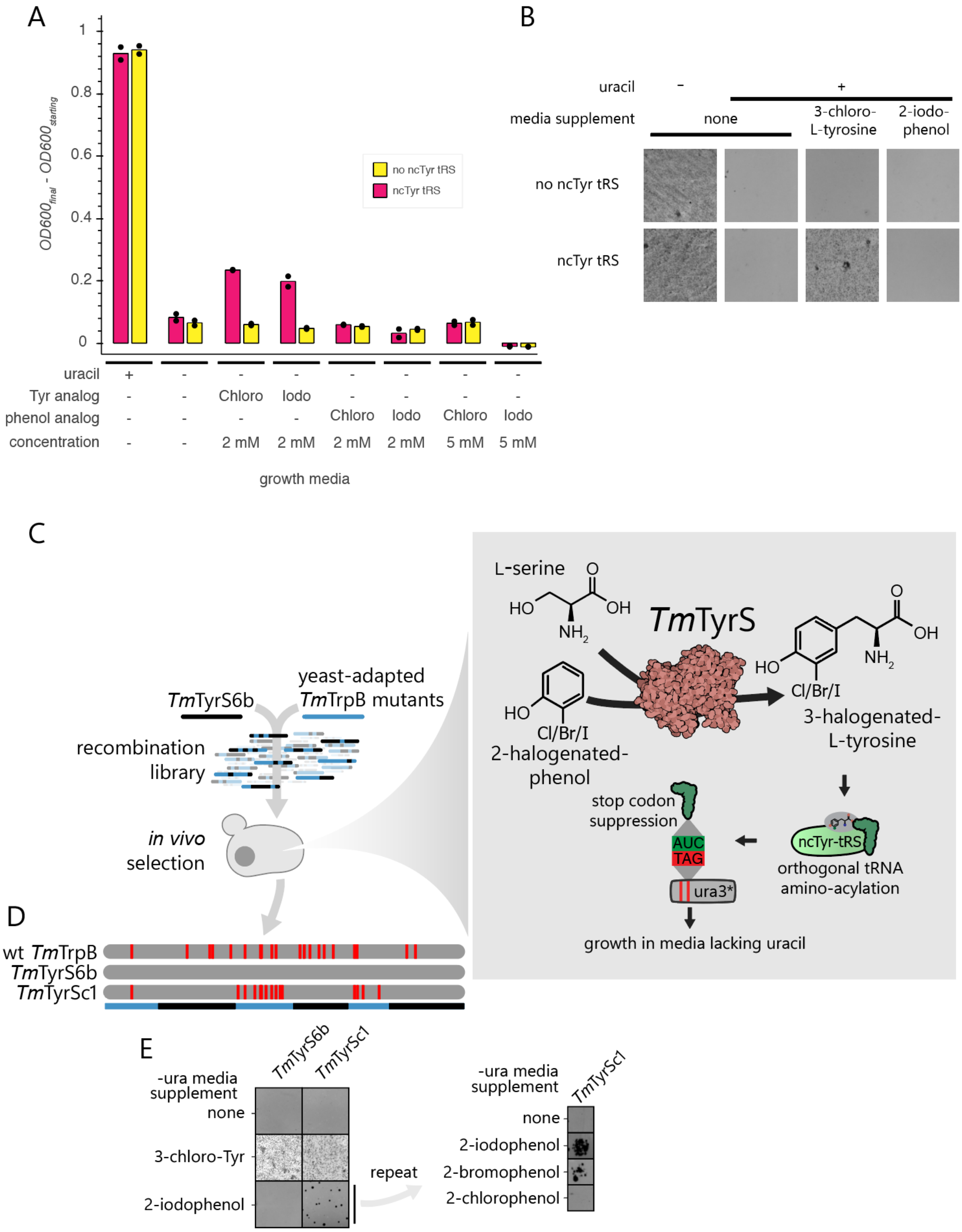

A
1
0.8
0.6
0.4
0.2
0
$OD600_{final}$ - $OD600_{starting}$
no ncTyr tRS
ncTyr tRS
uracil + - - - - - - -
Tyr analog - - Chloro Iodo - - - -
phenol analog - - - - Chloro Iodo Chloro Iodo
concentration - - 2 mM 2 mM 2 mM 2 mM 5 mM 5 mM
growth media
B
uracil - +
media supplement none 3-chloro-L-tyrosine 2-iodo-phenol
no ncTyr tRS
ncTyr tRS
C
TmTyrS6b
yeast-adapted TmTrpB mutants
recombination library
in vivo selection
L-serine
HO
OH
O
$NH_2$
TmTyrS
HO
Cl/Br/I
2-halogenated-phenol
HO
Cl/Br/I
OH
O
$NH_2$
3-halogenated-L-tyrosine
stop codon suppression
AUC
TAG
ura3*
ncTyr-tRS
orthogonal tRNA amino-acylation
growth in media lacking uracil
D
wt TmTrpB
TmTyrS6b
TmTyrSc1
E
-ura media supplement
TmTyrS6b
TmTyrSc1
none
3-chloro-Tyr
2-iodophenol
repeat
-ura media supplement
TmTyrSc1
none
2-iodophenol
2-bromophenol
2-chlorophenol

**Figure 2. Isolating ncTyrS variants with detectable *in vivo* activity via growth-based selection.** A) Validation of ncTyr tRNA synthetase (ncTyr tRS) by liquid growth assay. Yeast expressing a stop suppressor tRNA and either a corresponding ncTyr tRS or no additional tRS were grown in the indicated liquid media. Bars represent mean change in OD600 over a 48 hour growth period for n=2 biological replicates. B) Images of yeast expressing TmTyrS6b, a stop suppressor tRNA, and either a corresponding engineered ncTyr tRNA synthetase or no additional tRNA synthetase plated onto the indicated growth media and grown for 3 days. C) Illustration of the *in vitro* recombination and growth based selection used to isolate TyrSc1. D) Amino acid sequence alignment of wild type *Tm*TrpB and *Tm*TyrSc1, an isolate from the selection, to *Tm*TyrS6b. Predicted provenance of regions of *Tm*TyrSc1 based on this alignment are shown, indicating which regions are likely derived from yeast-adapted *Tm*TrpB mutants (blue) and which are likely derived from *Tm*TyrS6b (black). E) Images of yeast expressing *Tm*TyrS6b or *Tm*TyrSc1 plated onto growth media lacking uracil and supplemented with the indicated small molecule. Colonies exhibiting 2-iodophenol-dependent growth were harvested and were again plated onto the indicated growth media lacking uracil.

## An orthogonal tRNA / tRNA synthetase pair for selection for *in vivo* noncanonical tyrosine analog (ncTyr) production

*Tm*TrpB variants capable of producing various ncTyr analogs *in vitro* were recently identified through an extensive directed evolution campaign. The final variant identified in this campaign, deemed *Tm*TyrS6, is capable of producing multiple 3-halogenated-L-tyrosine species, namely 3-chloro-, 3-bromo-, and 3-iodo-L-tyrosine, with >60% yield[10]. To develop a growth-based selection that might be suitable for further improving these activities, we searched for engineered orthogonal tRNA / tRNA synthetase pairs capable of accepting 3-halogenated ncTyr analogs. This search returned a variant of the *Methanomethylophilus alvus* pyrrolysyl-tRNA synthetase, NitroY-RS-F5 (hereafter, ncTyr tRS), that was recently engineered to accept a range of ncAAs, among them 3-halogenated ncTyr analogs[11]. Furthermore, another group's work on using another variant of the *Methanomethylophilus alvus* tRS / tRNA pair suggested this pair was likely to be functional and orthogonal in yeast[12].

We installed *Tm*TyrS6 and this ncTyr tRS/tRNA pair into our ura3 stop codon suppression selection system and evaluated 3-chloro- and 3-iodo-L-tyrosine stop codon suppression and *in vivo* production. We found that, consistent with the reported ncTyr tRS substrate profile, 3-chloro- and 3-iodo-L-tyrosine were incorporated sufficiently to improve growth in media lacking uracil (Fig. 2A-B). However, despite the presence of *Tm*TyrS6, there was no indication of 2-chloro or 2-iodophenol-dependent survival (Fig. 2A-B).

## Domestication of a TrpB variant that produces Tyrosine analogs *in vivo*

We suspected that the long directed evolution history of *Tm*TyrS6 in which *in vitro*, but not *in vivo*, activity was screened for might have resulted in some amino acid residues that are suboptimal for *in vivo* activity in yeast. In contrast, our previous efforts in adapting *Tm*TrpB for activity in yeast was likely rich in broadly activating mutations, but of course lacked the hard-earned necessary mutations to alter substrate specificity toward phenol analogs. With this in mind, our efforts to adapt the enzyme to *in vivo* activity in yeast was two-fold. First, we consulted our library of deeply diversified TrpBs to identify candidate mutations to remove or add manually. We identified and removed four mutations that had been obtained prior to any screens for Tyr analog production but altered residues that were highly conserved within our library (E30G, I69V, K96L, L213P), three of which were easily accessible from frequent single nucleotide substitutions. We also added three mutations that were frequently found within our library (A20T, A118V, N167D). We call the resulting variant *Tm*TyrS6b. Second, we used our diverse library of *Tm*TrpB variants that had been selected for Trp production as a source of broadly activating mutations. Through *in vitro* recombination of *Tm*TyrS6b with indole-evolved *Tm*TrpBs harboring ~10-20 nonsynonymous mutations, we generated an 800-thousand member library of diverse variants.

We applied our growth-based selection to this library, using uracil-deficient solid growth media supplemented with 2-chloro-, 2-bromo-, or 2-iodo-phenol for positive selection (Fig. 2C). We also used a negative selection in which media supplemented with 5-fluoroorotic acid, which is converted into a toxic metabolite by URA3, is used to purge from the population cells that can produce full-length URA3 in the absense of any phenol analogs. Following two rounds of positive selection separated by one round of negative selection, we were able to isolate populations from both 2-iodophenol and 2-bromophenol selections that exhibited the desired phenol analog dependent uracil prototrophy phenotype (Fig. 2D-E). One representative sequence, which we call *Tm*TyrSc, differed from the parental *Tm*TyrS6b and from the wild type *Tm*TrpB sequences by 16 and 20 amino acid residues respectively. That *Tm*TyrSc had gained the desired phenotype despite this large number of mutations demonstrates a heretofore unappreciated use case for the large mutation-rich sequence libraries that OrthoRep is capable of generating.

*Tm*TyrSc may serve as a valuable starting point for further engineering. The selection we have developed here may be used in rounds of standard directed evolution or in continuous evolution using an *in vivo* hypermutation system such as OrthoRep. If it is used in a continuous format, this selection will require further development to ensure its longevity. For instance, using an AND gate with an additional amber-truncated auxotrophy marker may reduce the probability of genomic mutations that compromise the efficacy of the selection[13]. Alternatively, it may be possible use a tyrosine auxotroph strain to directly select for tyrosine production from exogenously provided phenol, as *Tm*TyrS6 has been shown to be capable of performing this reaction, albeit with quite low turnover frequency (~15 turnovers per hour)[10]. Our efforts thus far to use *Tm*TyrSc or *Tm*TyrS6 to detect this activity have thus far been unsuccessful (data not shown). The most promising path forward may therefore be to first improve upon the already detectable ncTyr production activity followed by evolving improvements to tyrosine production. If these efforts are successful, the resulting evolution campaign would represent a dramatic transformation for the substrate specificity of *Tm*TrpB, and may not only unlock efficient *in vivo* production of a diverse array of ncTyr species, but would represent the biosynthesis of an essential nutrient via a chemistry not observed in nature.

## 4.3 Materials and methods

### Plasmids and cloning

Plasmids and exact DNA sequences for relevant genes used in this study are listed in Tables 4.1 and 4.2, respectively. All plasmids were generated via Gibson assembly using PCR amplicons generated using Q5 Hot Start DNA Polymerase (NEB), followed by transformation into chemically competent *E. coli* strain TOP10 . All primers were purchased from IDT.

Important regions of plasmids were sequenced confirmed by Sanger sequencing (Azenta).

| Plasmid ID | Plasmid name | Use |
|---|---|---|
| pGR179 | pCAMCS-URA3.H26*.K75*-EcTrpRSh14-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR182 | pCAMCS-URA3.H26*.K75*-WT.EcTrpRS-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR183 | pCAMCS-URA3.H26*.K75*-noTrpRS-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR207 | pCAMCS-TmTriple-ura3.H26*.K75*-EcTrpRSh14-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR208 | pCAMCS-TmTriple-ura3.H26*.K75*-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR209 | pCAMCS-TmTriple-ura3.H26*.K75*-WTEcTrpRS-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR213 | pCAMCS-Q90*-003-1-A-ura3.H26*.K75*-EcTrpRSh14-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR214 | pCAMCS-Q90*-003-1-A-ura3.H26*.K75*-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR215 | pCAMCS-Q90*-003-1-A-ura3.H26*.K75*-WTEcTrpRS-ScArgtDNAUCU.EcTrptDNACUA-GFPTAG-HIS3 | Fig. 4.1B |
| pGR691 | pCAMCS-NitroYRS.F5(TDH3p)-URA3.H26*.K75*(TEF1p)-RxG(GAL1p)-HIS3 | Fig. 4.2 |
| pGR692 | pCAMCS-noRS-URA3.H26*.K75*(TEF1p)-RxG(GAL1p)-HIS3 | Fig. 4.2A |

**Table 4.1. Plasmids used in this study**

| gene | nucleotide sequence |
|---|---|
| ncTrp RS (EcTrp RS h14) | ATGACTAAGCCCATCGTTTTTgctGGCGCACAGCCCTCAGGTGAATTGACCATTGGTAACTACATGGGTGCGCTGCGTCAGTGGGTAAACATGCAGGATGACTACCATTGCATTTACTGTATCGTTGACCAACACGCGATCACCGTGCGCCAGGATGCACAGAAGCTGCGTAAAGCGACGCTGGATACGCTGGCCTTGTATCTGGCTTGTGGTATCGATCCTGAGAAAAGCACCATTTTTGTTCAGTCCCACGTGCCGGAACATGCACAGTTAGGCTGGGCACTGAACTGCTATACCTACTTCGGCGAACTGAGTCGCATGACGCAGTTTAAAGATAAATCTGCGCGTTATGCCGAGAACATCAACGCTGGTCTGTTTGACTATCCGGTGCTGATGGCAGCGGACATCCTGCTGTATCAAACTAATCTGggtccttgtGGTGAAGACCAGAAACAGCACCTCGAACTGAGCCGCGATATTGCCCAGCGTTTCAACGCGCTGTATGGCGAGATCTTTAAGGTGCCGGAGCCGTTTATTCCGAAATCTGGCGCGCGCGTAATGTCGCTGCTGGAGCCGACCAAGAAGATGTCCAAGTCTGACGATAATCGCAATAACGTTATCGGCCTGCTGGAAGATCCGAAATCGGTAGTGAAGAAAATCAAACGTGCGGTCACTGACTCCGACGAGCCGCCGGTAGTTCGCTACGATGTGCAGAACAAAGCGGGCGTTTCCAACCTGTTGGATATCCTTTCAGCGGTAACGGGCCAGAGCATCCCAGAACTGGAAAAACAGTTCGAAGGCAAGATGTATGGTCATCTGAAAGGTGAAGTGGCTGATGCCGTTTCCGGTATGCTGACTGAATTGCAGGAACGCTATCACCGTTTCCGCAACGATGAAGCCTTCCTGCAACAGGTGATGAAAGATGGCGCGGAAAAAGCCAGCGCGCACGCTTCCCGTACGCTAAAAGCGGTGTACGAAGCGATTGGTTTTGTGGCGAAGCCGTAA |
| ncTyr RS (MaPyl RS NitroY-F5) | ATGGCAGTAAAGTACACAGACGCTCAGATTCAGCGTCTGAGGGAGTACGGCAACGGTACTTACGAGCAAAAAGTCTTTGAAGACTTGGCTTCCAGGGATGCGGCTTTCTCAAAAGAAATGTCTGTAGCGTCCACCGATAACGAAAAGAAAATCAAGGGGATGATCGCGAATCCAAGTAGGCACGGTCTGACACAGTTGATGAACGATATCGCCGATGCATTGGTTGCAGAGGGCTTCATTGAGGTTAGAACGCCAATCTTCATATCTAAAGATGCTCTGGCCAGAATGACCATAACCGAAGATAAGCCACTTTTCAAGCAAGTGTTTTGGATTGACGAAAAGCGTGCGTTGAGGCCAATGTTAGCTCCGAACCTATATTCTGTGATGAGAGACTTGAGAGATCACACCGACGGACCAGTCAAGATTTTTGAGATGGGGAGCTGTTTCAGAAAGGAATCACACTCTGGAATGCATTTAGAAGAGTTTACCATGCTTGCTCTATCTGACATGGGTCCCAGAGGGGATGCAACGGAAGTTCTAAAGAACTACATCTCTGTAGTGATGAAAGCTGCTGGCTTGCCCGACTACGATCTGGTGCAGGAAGAGTCCGATGTTTATAAAGAGACAATCGACGTGGAAATCAACGGGCAAGAGGTTTGTTCAGCGTGTGTAGGTCCTCATTACTTGGACGCTGCGCATGATGTTCACGAACCTAGAAGCGGAGCAGGTTTTGGCCTAGAGAGGCTGCTGACTATACGTGAGAAATATTCAACGGTCAAGAAGGGCGGAGCGAGTATCTCCTACCTAAATGGAGCGAAAATTAAC |
| TmTriple | ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCAGAAATCCTGATGGGAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGAAATCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTTCTGCTGGCAAAAAAAATGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGTTCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACCGAACGTTGAACGTATGAAACTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTAACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGATCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCTGGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGGGTGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACTTACACGGTTCTAAGACGTTCGTTCTGCAGGATGACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTGTCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCACCACCACCACCACCACTGA |
| Q90*-003-1-A | ATGAAAGGCTACTTCGGTCCGCACGGTGGCCAGTACGTGCCAGAAATCCTGATGGGAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGAAATCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTACGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTTCTGCTGGCAAAAAAAACGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGTTCGGTATGGAATGTGTGATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACCGAATGTTGAACGTATGAAACTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTAATCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTGCGTAACTTCCAAAAGGTCATCGGTGAAGAGACCAAAAAACAGATTCTGGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGGGTGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATCCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACTTACACGGTTCTAAGACGTTCGTTCTGCAGGATGACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACTGCCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCACCATCACCATCACCATTAA |
| TmTyrS6 | ATGccagatAATTTTGGTCCATATGGTGGTCAAAATGTTCCTGAAATTTTGATGGGTGCTTTGGAAGAATTGGAAGCTGCTTATGAAGGTATTATGAAAGATGAATCTTTTTGGAAAGAATATAATGATTTGTTGAGAGATTATGCTGGTAGACCAACTCCATTGTATTTTGCTAGAAGATTGTCTGAAAAATATGGTGCTAGAGTTTATTTGAAAAGGGAAGATTTGTTGCATACTGGTGCTCATAAAATAAATAACACTATTGGTCAAGTTTTGTTGGCTAAATTGATGGGTAAAACTAGAATTACTGCTGGTACTGGTGCTGGTCAACATGGTGTTGCTACTGCTACTGCTGCTGCTTTGTTTGGTATGGAATGTGTTGTTTATATGGGTGAAGAAGATACTATTAGACAAAGATTGAATGTTGAAAGAATGAAACAATTGGGTGtTAAAGTTGTTCCTGTTAAATCTGGTTCTAGAACTTTGAAAGATGCTATTGATGAAGCTTTCAGAGATTGGACTACTAATTTGCAAACTACTCATTATGTTGCTGGTTCTGTTGTGGGACCATATCCATATCCTATAATTGTTAGAAACTTTCAGAAGGTTATTGGTGAAGAAACTAAGAAACAAATTCCTGAAAAAGAAGGTAGATTGCCTGATTATATTGTTGCTTGTATGTCTGcTGGTTCTAATGCTGCTGGTATATTTTATCCATTTATTGATTCTGGTGTTAAATTGATTGGTGTTGAAGCTGGTGGTGAAGGTTTGGAAACTGGTAAACATGCGGCTTCTTTGTTGAAAGGTAAAATTGGTTATTTGCATGGTTCTAAAACTTTTGTTTTGCAAGATGATGGTGGTCAAGTTCAAGCTTCTCATTCTGTTTCTGCTGGTTTGGATTATCCTGGTGTTGGTCCTGAACATGCTTATTGGAGAGAGACGGGTAAAGTCCTTTATGATGCTGTTACTGATGAAGAAGCTTTGGATGCTTTTATTGAATTGTCTAGATTGGAAGGTATTATTCCTGCTTTGGAATCCTCTCATGCTTTGGCATACTTGAAGAAGATTAATATTAAGGGTAAAGTTGTTGTTGTTAATTTGTCTGGTAGAGGTGATAAAGATTTGGAATCTGTTTTGAATCATCCATATGTTAGAGAAAGAATTCATTTGGAAcacCATcacCATcacCATtga |
| TmTyrS6b | ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGCCAACTCTGGAAGAACTGGAAGCTGCGTACGAAGAAATCATGAAAGATGAGTCTTTCTGGAAAGAATTTAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAgAAATACGGTGCTCGCATCTATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAAgAAgATGGGCAAgACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGTTGCAGCGCTGTTCGGTATGGAATGTGTAGTTTATATGGGCGAAGAAGACACGATCCGCCAGAGGTTGAACGTTGAACGTATGAAACAACTGGGTGTTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGATGAAGCTTTCCGTGACTGGACTACCAACCTGCAGACCACCCATTACGTGGCTGGCTCTGTGGTTGGTCCGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAAACCAAgAAgCAGATTCTGGAgAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCATGTCTGCTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACGGTGGTCAAGTTCAGGTGTCTCACTCCGTTTCCGCTGGCCTGGACTACTCCGGTGTCGGTCCGGAACACGCCTATTGGCGTGAGACTGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCATCATCATCATCATCATTAA |
| TmTyrSc | ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGCCAGTTCTGGAAGAACTGGAAGCTGCGTACGAAGAAATCATGAAAGATGAGTCTTTCTGGAAAGAATTTAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAGAAATACGGTGCTCGCATCTATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAAGAAGATGGGCAAGACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGTAGCAGCGCTGTTCGACATGGAATGTGTAATCTATATGGGCGAAGAAGACATGATCCGCCAGAAACCGAACGTTAAACGTATGAAACTACTGGGTGCTAAAGTTATAACGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGATGAAGCTTTCCGTGACTGGACTACCAACCTGCAGACCACCCATTACGTGGCTGGCTCTGTGGTTGGTCCGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAAACCAAGAAGCAGATTCTGGAGAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGGGTGGTGGTTCTAATGCTGTCGGTATCTTCTATCCGTTTATCGATTCTGGTGCGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCACTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACGGTGGTCAAGTTCAGGTGTCTCACTCCGTTTCCGCTGGCCTGGACTACTCCGGTGTCGGTCCGGAACACGCCTATTGGCGTGAGACTGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCATCATCATCATCATCATTAA |

**Table 4.2. Complete DNA sequences for genes used in this study.**

### Yeast strains, media, transformations, and DNA extraction

All yeast strains used in this study were derived from *S. cerevisiae* AH22 (ATCC ID# 38626), with the entire ORFs for either TRP5 (Fig. 4.1) or both TRP5 and TYR1 (Fig. 4.2) knocked out. Yeast were grown in liquid or on plates at 30 °C in synthetic complete (SC) growth medium (20 g/L dextrose, 6.7 g/L yeast nitrogen base w/ ammonium sulfate w/o amino acids (US Biological), appropriate nutrient drop-out mix (US Biological), as directed).

Yeast transformations, including p1 integrations, were performed as described in Gietz and Shiestl. Briefly, a 4 mL culture of yeast was grown to mid-log phase in rich YPD medium (2% (w/v) Bacto yeast extract (US Biological), 4% (w/v) Bacto peptone (US Biological), 4% (w/v) glucose) at 30 °C, harvested by centrifugation, washed with sterile water, and pelleted again by centrifugation. These pellets were then resuspended in a mixture containing PEG3350 (Fisher) (30% (w/v) final concentration), lithium acetate (Sigma Aldrich) (90 mM final concentration), boiled salmon sperm carrier DNA (0.25 mg/mL final concentration), and ~1 µg of the DNA to be transformed, all in a total volume of 410 µL.

Extraction of genomic DNA (gDNA) and p1/p2 plasmids was performed as previously described (see Chapter 1, Materials and Methods).

### Plating assays, liquid growth assays, and TyrS selection

Spot plating assays (*e.g.* Fig. 1B) were performed by growing the appropriate yeast strain in media selective for any plasmids, diluting the saturated yeast culture 1:100 into 0.9% NaCl, and plating 10 µL of this dilution onto the appropriate plating media. Plating assays using a larger plate surface area (*e.g.* Fig. 2A) were used in cases where only a low fraction of cells exhibit the desired phenotype, and were performed by plating 100 µL of saturated yeast culture onto solid media in a 10 cm wide petri dish. Liquid growth assays were performed by passaging 2 µL into 200 µL of liquid culture, then measuring OD600 immediately and after 24 hours of growth at 30 °C with shaking. The difference in OD600 between the two timepoints is reported. Positive and negative selection of the recombination library were both performed on solid media. Plating assays and positive selections for 2-chloro, 2-bromo, and 2-iodophenol were performed at 1, 0.5, and 0.5 mM concentrations, respectively. Negative selections were performed with 1 g/L 5FOA without phenol analog.

## 4.4 References

1. Plass, T., Milles, S., Koehler, C., Schultz, C. & Lemke, E. A. Genetically encoded copper-free click chemistry. *Angew. Chemie - Int. Ed.* **50**, 3878–3881 (2011).

2. Wu, N., Deiters, A., Cropp, T. A., King, D. & Schultz, P. G. A genetically encoded photocaged amino acid. *J. Am. Chem. Soc.* **126**, 14306–14307 (2004).

3. Burke, A. J. *et al.* Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature.* **570**, 219–223 (2019).

4. Almhjell, P. J., Boville, C. E. & Arnold, F. H. Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997 (2018).

5. Biava, H. D. Tackling Achilles' Heel in Synthetic Biology: Pairing Intracellular Synthesis of Noncanonical Amino Acids with Genetic-Code Expansion to Foster Biotechnological Applications. *ChemBioChem* **21**, 1265–1273 (2020).

6. Zeymer, C. & Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **87**, 131–157 (2018).

7. Italia, J. S. *et al.* An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat. Chem. Biol.* **13**, 446–450 (2017).

8. Hancock, S. M., Uprety, R., Deiters, A., Chin, J. W. & Carolina, N. Expanding the Genetic Code of Yeast for Incorporation of Diverse Unnatural Amino Acids via a Pyrrolysyl-tRNA Synthetase / tRNA Pair. 14819–14824 (2010).

9. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A Panel of TrpB Biocatalysts Derived from Tryptophan Synthase through the Transfer of Mutations that Mimic Allosteric Activation. *Angew. Chemie - Int. Ed.* **55**, 11577–11581 (2016).

10. Almhjell, P. J. Noncanonical Amino Acid Synthesis by Evolved Tryptophan Synthases Thesis by. **2023**, (2023).

11. Avila-Crump, S. *et al.* Generating Efficient Methanomethylophilus alvus Pyrrolysyl-tRNA Synthetases for Structurally Diverse Non-Canonical Amino Acids. *ACS Chem. Biol.* **17**, 3458–3469 (2022).

12. Stieglitz, J. T., Lahiri, P., Stout, M. I. & Van Deventer, J. A. Exploration of Methanomethylophilus alvus Pyrrolysyl-tRNA Synthetase Activity in Yeast. *ACS Synth. Biol.* (2022) doi:10.1021/acssynbio.2c00001.

13. Tan, L. *et al.* Efficient Selection Scheme for Incorporating Noncanonical Amino Acids Into Proteins in Saccharomyces cerevisiae. *Front. Bioeng. Biotechnol.* **8**, 1–10 (2020).

# Chapter 5. Summary and future directions

## 5.1 TrpB evolution

Our application of OrthoRep to TrpB evolution in this work demonstrates the ability of OrthoRep to generate immense sequence diversity as well as novel ways to utilize this diversity, having used it for mining promiscuous functions of synthetic orthologs, studying molecular evolution, and improving upon noncognate TrpB activities. There are also many exciting new directions to pursue. For instance, given that our work focused almost exclusively on studying TrpB evolution, the generalizability of these approaches should be investigated. More specifically, can OrthoRep serve as a more general platform for generating and mining useful biomolecular diversity? The ease with which OrthoRep can be scaled coupled with the computational methods that we have developed means that the answer to this question is well within reach. Through selection for genes encoded on the OrthoRep system over relatively short timescales (1-2 months) and deep, highly accurate, long read sequencing, it is now possible to generate datasets comprising hundreds of thousands of highly diverse variants of these genes, rivaling the diversity of orthologous genes in nature. But unlike natural sequence datasets, these sequences represent many independent examples of evolution occuring under virtually identical conditions. Using sequence barcodes that enable separation of distinct evolutionary lineages occuring even in the same physical population lends even further scalability. Such experimental approaches may be key in gaining a better understanding of broadly applicable principles of molecular evolution. Application of these approaches to orphan or *de novo* proteins, for which natural diversity is sparse or nonexistent, may also advance our understanding of poorly understood or novel functions and how they are shaped by evolution.

These efforts will benefit greatly from expanding the functions that can be selected for using OrthoRep. In Chapter 4, we describe our efforts in establishing a direct selection for ncAA biosynthesis using engineered tRNA synthetases as ncAA biosensors. We successfully engineered minor improvements to the *in vivo* function of the previously engineered TrpB mutant *Tm*TyrS6 using this approach, but more work needs to be done before it can be efficiently used in combination with OrthoRep. Tunable selection strength, which can precipitate large improvements in gene fitness during continuous evolution[1,2], would be a valuable feature. Tunable GOI expression is the most broadly applicable method of implementing selection tuning, but would likely require engineering a new system for cytoplasmic expression in yeast (see Section 5.2). Alternatively, implementing established systems for tuning expression of nuclear-encoded genes in yeast[3] could enable tunable selection strength, for instance through inducible expression of the tRNA synthetase or amber-truncated auxotrophy markers. Furthermore, the robustness of the selection over long evolution campaigns remains untested, but is likely to be compromised by genomic mutations that disrupt the selection. It may therefore improve robustness to implement an AND gate selection architecture, in which more than one phenotype rendered dependent upon *in vivo* ncAA production are is selected for.

The original engineering efforts that resulted in the *Tm*TyrS6 TrpB mutant demonstrated that, while many L-tyrosine analogs could be produced *in vitro* at high yields, only trace amounts of Tyr itself could be produced. Despite significant efforts to do so, we have not yet been able to detect *in vivo* Tyr production from *Tm*TyrS (data not shown). Our selection may provide a route for improving upon this activity via initial improvements to ncTyr production, followed by direct selection for Tyr production using a Tyr auxotroph. This would yield a novel biosynthetic pathway for an essential nutrient and, if performed at scale using OrthoRep, could provide valuable datasets for studying dramatic evolutionary transformations on laboratory timescales.

## 5.2 OrthoRep development

In this work we presented our successful efforts to engineer the workhorse of the OrthoRep system—TP-DNAP1—to exhibit a mutation rate of $10^{-4}$ substitutions per base per generation, 10-fold higher than previous TP-DNAP1 variants. When coupled with the robustness of the direct hypermutation system (see Chapter 1) in OrthoRep, these engineered polymerases enabled the sustained diversification of TrpB over hundreds of generations without any indication of decaying error rates. This new regime of *in vivo* hypermutation should have broad utility for the already diverse continuous evolution applications thus far developed. The dual selection- and screening- based approach we used to engineer these polymerases may also be applicable to further increasing mutation rates for OrthoRep or for nascent systems with the direct hypermutation architecture.[4]

While OrthoRep is now clearly capable of evolving thousands of highly diverse, functional sequences, the high fraction of nonfunctional sequences we observed in this work suggests increasing mutation rates is no longer the priority for further developing the OrthoRep platform. If individual sequences cannot reliably perform the desired function, it is likely that the speed of purifying selection could be improved. Generating large datasets of functional sequences under such conditions also requires laborious additional cloning and selection steps. We posit that the multi-copy nature of the OrthoRep plasmid and constitutive mutagenesis play a key role in this phenomenon, as each individual cell in principle only needs a single functional copy of the GOI making additional functional copies superfluous. Engineering systems for copy number control and mutagenesis regulation may therefore be a key next step in improving OrthoRep performance. Additionally, RNA transcription of GOIs on OrthoRep is currently accomplished by the native cytoplasmic transcription system of the natural plasmid system from which OrthoRep is derived, which does not support high levels of expression. Establishing alternative methods of cytoplasmic expression may enable higher expression levels, thereby increasing the minimum threshold of GOI activity that can be selected for. The complexity of these and other aspects of this replication system may merit pursuing a bottom-up approach to further engineering of OrthoRep.[5–7]

While these future developments should further improve the value of the OrthoRep platform, the work presented here represents a considerable technological advancement for *in vivo* hypermutation systems, a demonstration of their value for biomolecular engineering and an insightful study of rapid molecular evolution.

## 5.3 References

1. Carlson, J. C., Badran, A. H., Guggiana-Nilo, D. A. & Liu, D. R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat. Chem. Biol.* **10**, 216–222 (2014).
2. DeBenedictis, E. A. *et al.* Systematic molecular evolution enables robust biomolecule discovery. *Nat. Methods.* **19**, 55–64 (2022).
3. Bragdon, M. D. J. *et al.* Cooperative assembly confers regulatory specificity and long-term genetic circuit stability. *bioRxiv* **186**, 2022.05.22.492993 (2022).
4. Tian, R. *et al.* Engineered bacterial orthogonal DNA replication system for continuous evolution. *Nat. Chem. Biol.* (2023) doi:10.1038/s41589-023-01387-2.
5. Khalil, A. S. & Collins, J. J. Synthetic biology: Applications come of age. *Nat. Rev. Genet.* **11**, 367–379 (2010).
6. Yi, X., Khey, J., Kazlauskas, R. J. & Travisano, M. Plasmid hypermutation using a targeted artificial DNA replisome. *Sci. Adv.* **7**, (2021).
7. An, W. & Chin, J. W. Synthesis of orthogonal transcriptiontranslation networks. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 8477–8482 (2009).

# Appendix A: Supplementary information for Chapter 2

**Table S2.1 | Summary of all cultures passaged for evolution of *Tm*TrpB variants.**

| TrpB variant (strain) | culture volume (mL) | total number passaged | total number successful |
|---|---|---|---|
| wt *Tm*TrpB (GR-Y053) | 3 | 4 | 1 |
| | **100** | **2** | **2** |
| *Tm*Triple (GR-Y055) | 3 | 8 | 2 |
| | **100** | **4** | **4** |
| *Tm*TripleQ90* (GR-Y057) | 3 | 8 | 1 |
| | **100** | **0** | **0** |

**Table S2.2 | Mutation summary statistics for OrthoRep-evolved TrpB populations.**

| | | **non-synonymous mutations** | | | **synonymous mutations** | |
|---|---|---|---|---|---|---|
| variant set | **total number of sequences** | **total unique mutations** | **mean** | **standard deviation** | **mean** | **standard deviation** |
| consensus | **10** | **43** | **5.6** | **2.3** | **3.0** | **2.3** |
| 1 | 16 | 85 | 8.7 | 2.1 | 6.3 | 4.4 |
| 2 | **60** | **194** | **9.3** | **2.8** | **6.5** | **3.0** |

**Table S2.3 | Mutations and identification information for all individual *Tm*TrpB sequences.** Note that variant names are abbreviations of the evolution experiment from which the variant was harvested. For example, the variant name WT-100-1-A refers to the arbitrarily designated unique clone A that came from replicate 1 of the 100 mL evolution experiment that started from wt *Tm*TrpB. Likewise, Q90*-003-1-B refers to the arbitrarily designated unique clone B that came from replicate 1 of the 3 mL evolution experiment that started from *Tm*TripleQ90*. Likewise, Tri-003-2-A refers to the arbitrarily designated unique clone A that came from replicate 2 of the 3 mL evolution experiment that started from *Tm*Triple.

| variant set | variant name | number of non-synonymous mutations | number of synonymous mutations | starting sequence |
|---|---|---|---|---|
| 1 | WT-100-1-A | 13 | 3 | TmTrpB |
| **1** | **WT-100-2-A** | **9** | **7** | **TmTrpB** |
| 1 | WT-003-1-A | 7 | 1 | TmTrpB |
| **1** | **Q90*-003-1-A** | **6** | **11** | **TmTripleQ90*** |
| 1 | Q90*-003-1-B | 11 | 15 | TmTripleQ90* |
| **1** | **Tri-003-1-A** | **9** | **3** | **TmTriple** |
| 1 | Tri-003-1-B | 11 | 4 | TmTriple |
| **1** | **Tri-003-1-C** | **9** | **3** | **TmTriple** |
| 1 | Tri-003-2-A | 9 | 6 | TmTriple |
| **1** | **Tri-100-1-A** | **10** | **5** | **TmTriple** |
| 1 | Tri-100-2-A | 9 | 7 | TmTriple |
| **1** | **Tri-100-2-B** | **8** | **6** | **TmTriple** |
| 1 | Tri-100-3-A | 5 | 3 | TmTriple |
| **1** | **Tri-100-4-A** | **10** | **1** | **TmTriple** |
| 1 | Tri-100-4-B | 6 | 14 | TmTriple |
| **1** | **Tri-100-4-C** | **7** | **11** | **TmTriple** |
| 2 | WT-100-1-B | 14 | 9 | WT TmTrpB |
| **2** | **WT-100-1-C** | **12** | **7** | **WT TmTrpB** |
| 2 | WT-100-1-D | 12 | 9 | WT TmTrpB |
| **2** | **WT-100-1-E** | **14** | **10** | **WT TmTrpB** |
| 2 | WT-100-1-F | 16 | 11 | WT TmTrpB |
| **2** | **WT-100-2-B** | **7** | **9** | **WT TmTrpB** |
| 2 | WT-100-2-C | 7 | 5 | WT TmTrpB |
| **2** | **WT-100-2-D** | **9** | **8** | **WT TmTrpB** |
| 2 | WT-100-2-E | 8 | 9 | WT TmTrpB |

| | | | | |
|---|---|---|---|---|
| **2** | **WT-100-2-F** | **9** | **10** | **WT TmTrpB** |
| 2 | WT-100-2-G | 7 | 6 | WT TmTrpB |
| **2** | **WT-100-2-H** | **9** | **9** | **WT TmTrpB** |
| 2 | WT-100-2-I | 10 | 5 | WT TmTrpB |
| **2** | **WT-003-1-B** | **6** | **2** | **WT TmTrpB** |
| 2 | WT-003-1-C | 7 | 5 | WT TmTrpB |
| **2** | **WT-003-1-D** | **7** | **7** | **WT TmTrpB** |
| 2 | WT-003-1-E | 10 | 8 | WT TmTrpB |
| **2** | **WT-003-1-F** | **8** | **4** | **WT TmTrpB** |
| 2 | WT-003-1-G | 9 | 4 | WT TmTrpB |
| **2** | **WT-003-1-H** | **8** | **4** | **WT TmTrpB** |
| 2 | WT-003-1-I | 8 | 6 | WT TmTrpB |
| **2** | **Q90*-003-1-C** | **10** | **6** | **TmTripleQ90*** |
| 2 | Q90*-003-1-D | 11 | 9 | TmTripleQ90* |
| **2** | **Q90*-003-1-E** | **14** | **8** | **TmTripleQ90*** |
| 2 | Q90*-003-1-F | 14 | 8 | TmTripleQ90* |
| **2** | **Q90*-003-1-G** | **16** | **6** | **TmTripleQ90*** |
| 2 | Q90*-003-1-H | 12 | 9 | TmTripleQ90* |
| **2** | **Tri-003-1-D** | **9** | **4** | **TmTriple** |
| 2 | Tri-003-1-E | 7 | 5 | TmTriple |
| **2** | **Tri-003-1-F** | **12** | **3** | **TmTriple** |
| 2 | Tri-003-1-G | 10 | 5 | TmTriple |
| **2** | **Tri-003-1-H** | **15** | **6** | **TmTriple** |
| 2 | Tri-003-1-I | 6 | 4 | TmTriple |
| **2** | **Tri-003-2-B** | **8** | **11** | **TmTriple** |
| 2 | Tri-003-2-C | 13 | 4 | TmTriple |
| **2** | **Tri-003-2-D** | **12** | **8** | **TmTriple** |
| 2 | Tri-003-2-E | 10 | 3 | TmTriple |
| **2** | **Tri-100-1-B** | **5** | **5** | **TmTriple** |
| 2 | Tri-100-1-C | 6 | 5 | TmTriple |
| **2** | **Tri-100-1-D** | **9** | **6** | **TmTriple** |
| 2 | Tri-100-1-E | 10 | 8 | TmTriple |
| **2** | **Tri-100-1-F** | **7** | **2** | **TmTriple** |
| 2 | Tri-100-1-G | 6 | 5 | TmTriple |

| | | | | |
|---|---|---|---|---|
| **2** | **Tri-100-1-H** | **8** | **4** | **TmTriple** |
| 2 | Tri-100-2-C | 8 | 5 | TmTriple |
| **2** | **Tri-100-2-D** | **10** | **15** | **TmTriple** |
| 2 | Tri-100-2-E | 13 | 10 | TmTriple |
| **2** | **Tri-100-2-F** | **9** | **10** | **TmTriple** |
| 2 | Tri-100-2-G | 9 | 15 | TmTriple |
| **2** | **Tri-100-2-H** | **10** | **10** | **TmTriple** |
| 2 | Tri-100-2-I | 13 | 4 | TmTriple |
| **2** | **Tri-100-3-B** | **7** | **3** | **TmTriple** |
| 2 | Tri-100-3-C | 6 | 3 | TmTriple |
| **2** | **Tri-100-3-D** | **8** | **4** | **TmTriple** |
| 2 | Tri-100-3-E | 6 | 2 | TmTriple |
| **2** | **Tri-100-3-F** | **5** | **3** | **TmTriple** |
| 2 | Tri-100-4-D | 8 | 8 | TmTriple |
| **2** | **Tri-100-4-E** | **7** | **7** | **TmTriple** |
| 2 | Tri-100-4-F | 6 | 6 | TmTriple |
| **2** | **Tri-100-4-G** | **6** | **5** | **TmTriple** |

**Table S2.4 | Kinetic parameters of selected *Tm*TrpB variants at 30 °C.**

| variant | $k_{cat}$ [95% credible region] ($s^{-1}$) | $K_M$ [95% credible region] (μM) | $k_{cat}/K_M$ [95% credible region] ($mM^{-1}$ $s^{-1}$) |
|---|---|---|---|
| *Tm*Triple | 0.2 [0.16, 0.31] | 41.23 [14.32, 192.66] | 4.89 [1.54, 12.12] |
| **WT-003-1-A** | **0.53 [0.49, 0.58]** | **3.89 [1.85, 7.99]** | **137.22 [70.29, 276.04]** |
| Q90*-003-1-A | 0.77 [0.72, 0.83] | 5.79 [3.82, 8.8] | 133.38 [91.24, 193.71] |
| **Tri-100-2-A** | **0.62 [0.59, 0.66]** | **5.58 [3.99, 7.91]** | **111.89 [81.52, 152.25]** |

## Supplementary Figures

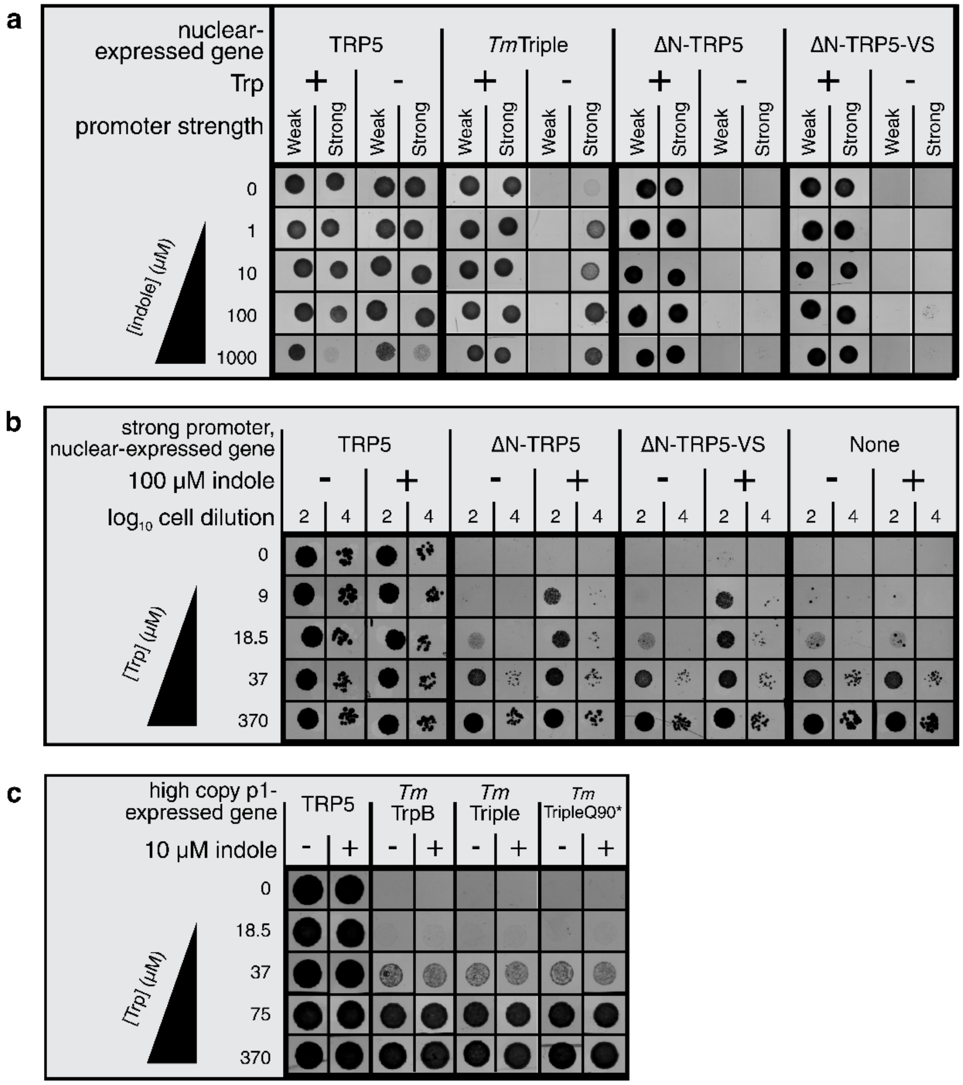


**Figure S2.1 | Evaluation of indole-dependent TRP5 complementation of TrpB variants. a-c**, Spot plating assays for Δ*trp5* yeast strains expressing TrpB variants from a nuclear plasmid under two different promoter strengths (**a**), from a nuclear plasmid under a strong promoter (**b**), or from p1 at a high copy number (wt TP-DNAP1 expressed *in trans*) (**c**) grown on indicated growth medium. ΔN-TRP5, N-terminally truncated yeast TRP5 constituting only the region of TRP5 homologous to *Tm*TrpB. ΔN-TRP5-VS, ΔN-TRP5 with two of the three *Tm*Triple mutations relative to wt *Tm*TrpB. The markedly reduced growth at 1000 μM indole only when TRP5 is expressed by a strong promoter may be explained by indole toxicity induced by additional indole production by TRP5.

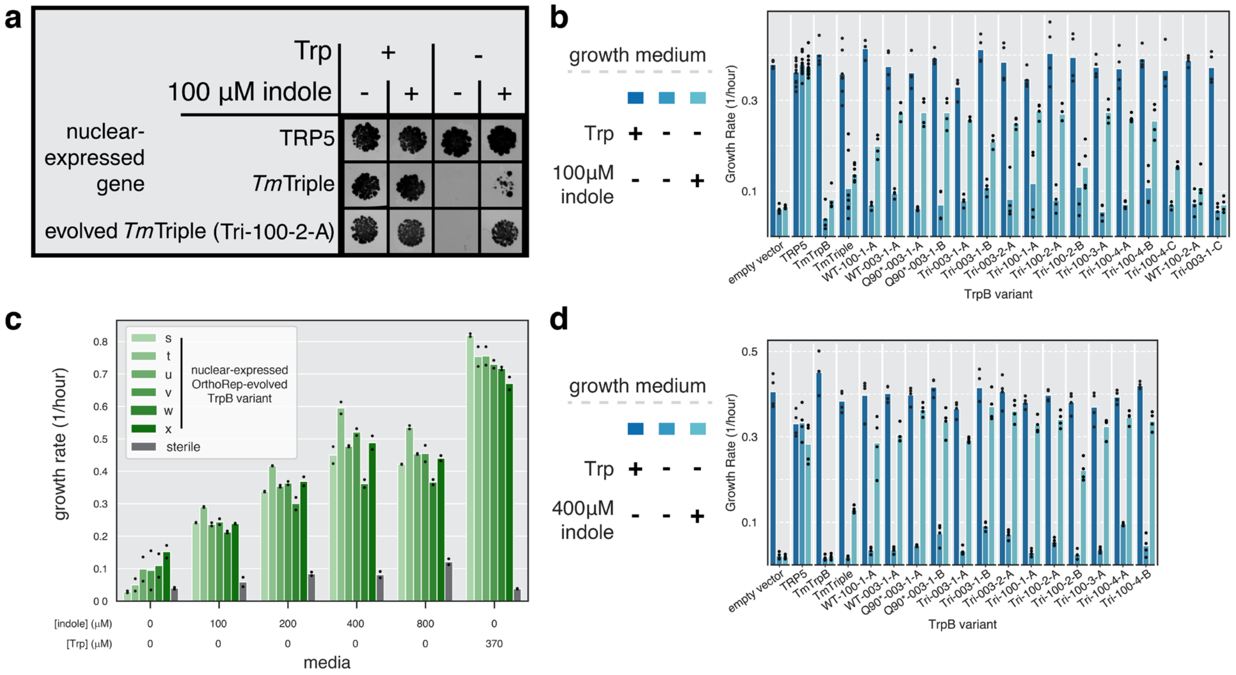


**Figure S2.2 | *In vivo* Trp production by evolved TrpBs. a**, Spot plating assay for TRP5-deleted yeast with a nuclear plasmid expressing TRP5, *Tm*Triple, or an individual OrthoRep-evolved *Tm*TrpB variant driven by a promoter (pRNR2) that approximates expression of *Tm*TrpB variants from OrthoRep's p1 plasmid, grown on indicated media. **b-d**, Evaluation of TRP5 complementation by evolved variants through a growth rate assay. Maximum growth rates during exponential growth phase (when rate is above ~0.15 per hour) over a 24-hour period for *Δtrp5* yeast strains transformed with a nuclear plasmid expressing the indicated *Tm*TrpB variant, grown in the indicated growth medium. Points represent growth rate for individual replicates, bars represent the mean growth rate for all replicates**.** Note that growth rates below ~0.15 per hour correspond to cultures that did not enter exponential phase; in these cases, the reported growth rate is not meaningful and instead can be interpreted as no quantifiable growth. **b,** Low indole growth rate test. All conditions tested in at least biological quadruplicate. **c**, Optimization of indole concentration. All conditions tested in technical duplicate, although exact sequences of *Tm*TrpB variants were not determined, as the sole purpose of these variants was to evaluate the effect of indole concentration. **d**, High indole growth rate test. All conditions tested in at least biological quadruplicate. A subset of this data is shown in **Fig. 2a**.

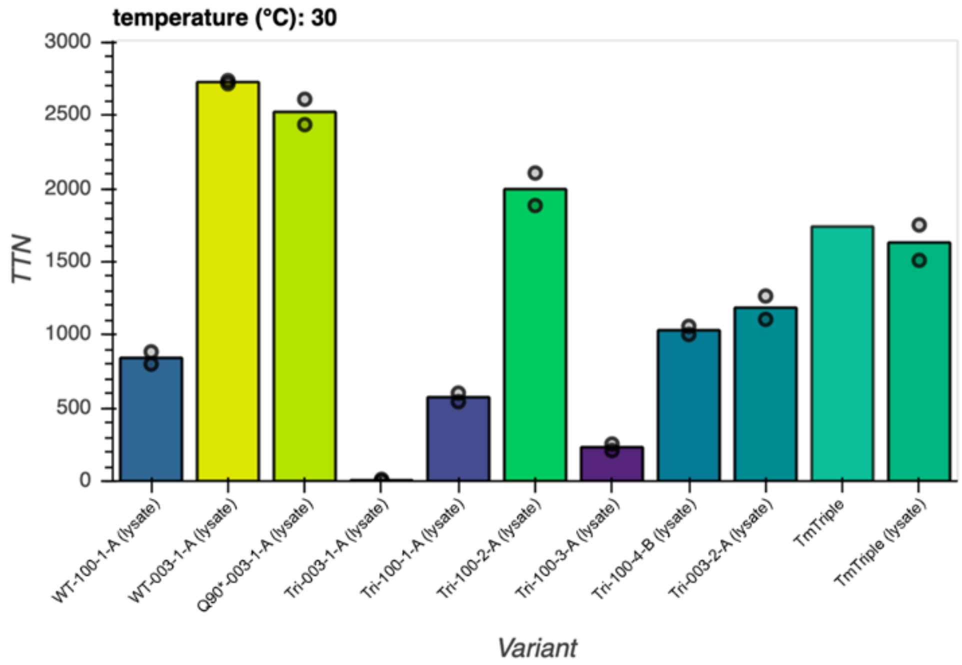


**Figure S2.3 | *In vitro* Trp production by evolved TrpBs with heat treated lysate.** Trp production at 30 °C by indicated *Tm*TrpB variants. Reactions with *Tm*Triple were performed with both heat treated lysate (1 hour incubation at 75 °C) and with purified protein, while all other reactions were performed only with heat treated lysate. TTN, total turnover number. Maximum TTN is 5,000. Points represent TTN for individual biological replicates, bars represent mean TTN for reactions with replicates, or TTN for a single replicate otherwise.

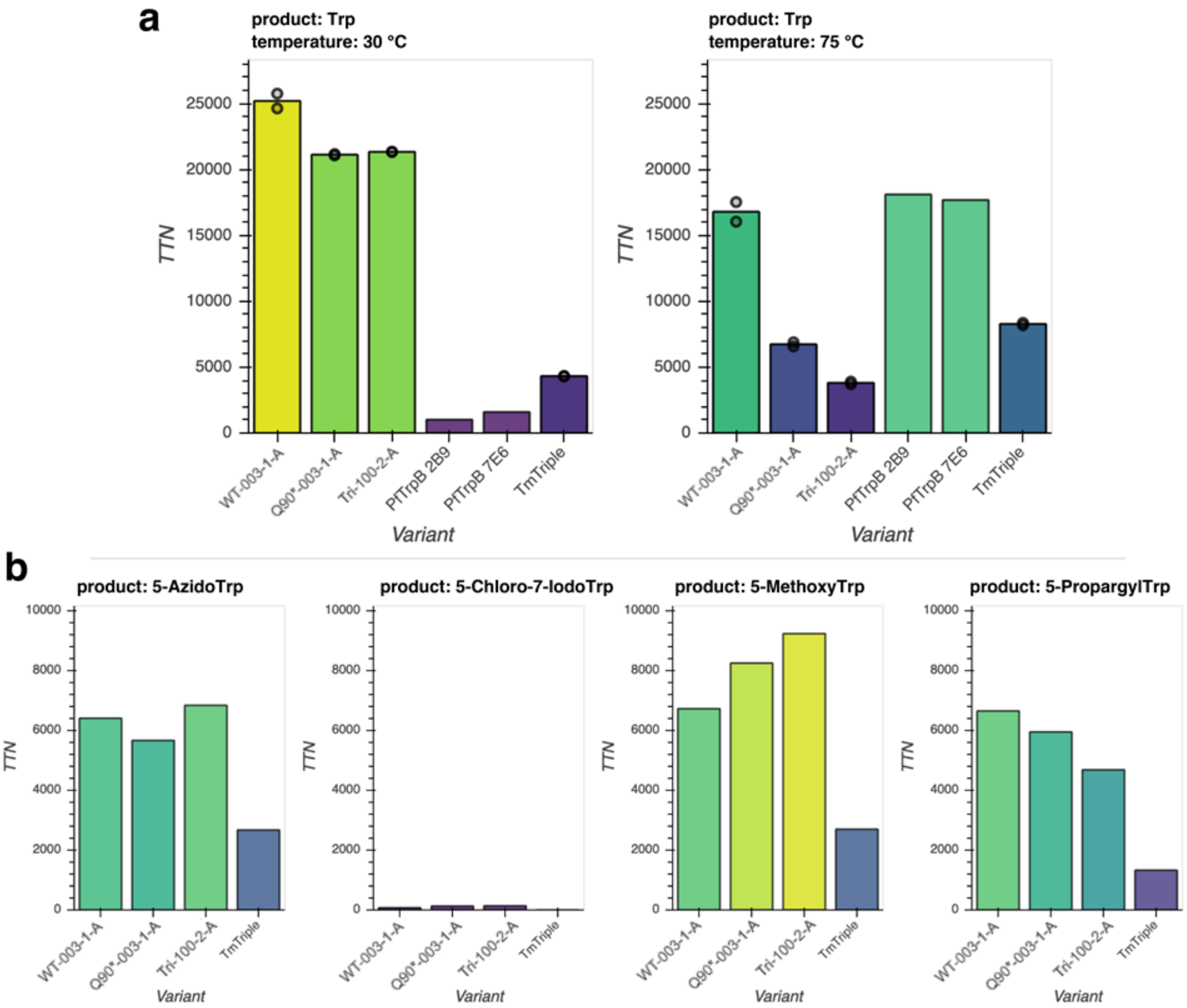


**Figure S2.4 | *In vitro* Trp and Trp analog production with purified enzyme. a**,**b**, Production of (**a**) Trp at 30 °C or 75 °C with 40,000 maximum TTN or (**b**) indicated Trp analogs at 30 °C by column purified *Tm*TrpB variants. TTN, total turnover number with 10,000 as maximum TTN. Points represent TTN for individual biological replicates, bars represent mean TTN for reactions with replicates, or TTN for a single replicate otherwise.

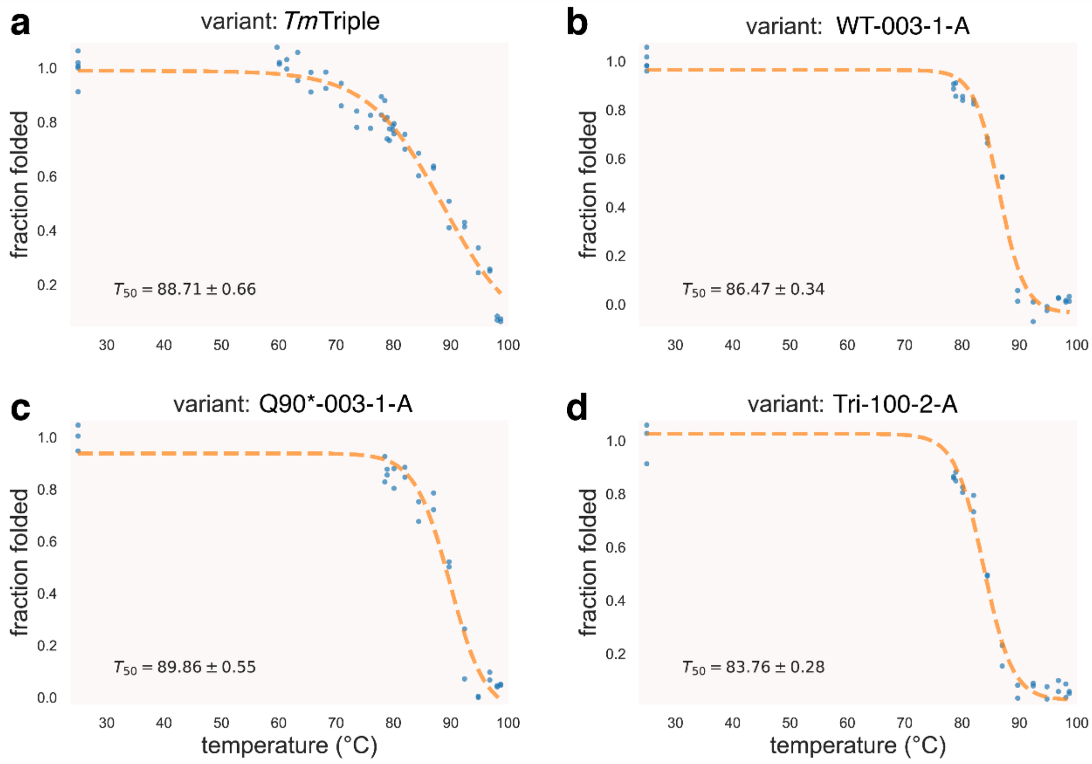


**Figure S2.5 | Thermal shift assay on various *Tm*TrpBs.** Proportion of *Tm*TrpB variants TmTriple (**a**), WT-003-1-A (**b**), Q90*-003-1-A (**c**), or Tri-100-2-A (**d**) that remain folded after incubation at indicated temperature for 1 hour, as measured by the fraction of Trp production relative to incubation at 25 °C. $T_{50}$, temperature at which 50% of enzyme is irreversibly inactivated, as estimated by best fit logistic model (dotted line). Each temperature tested in technical duplicate.

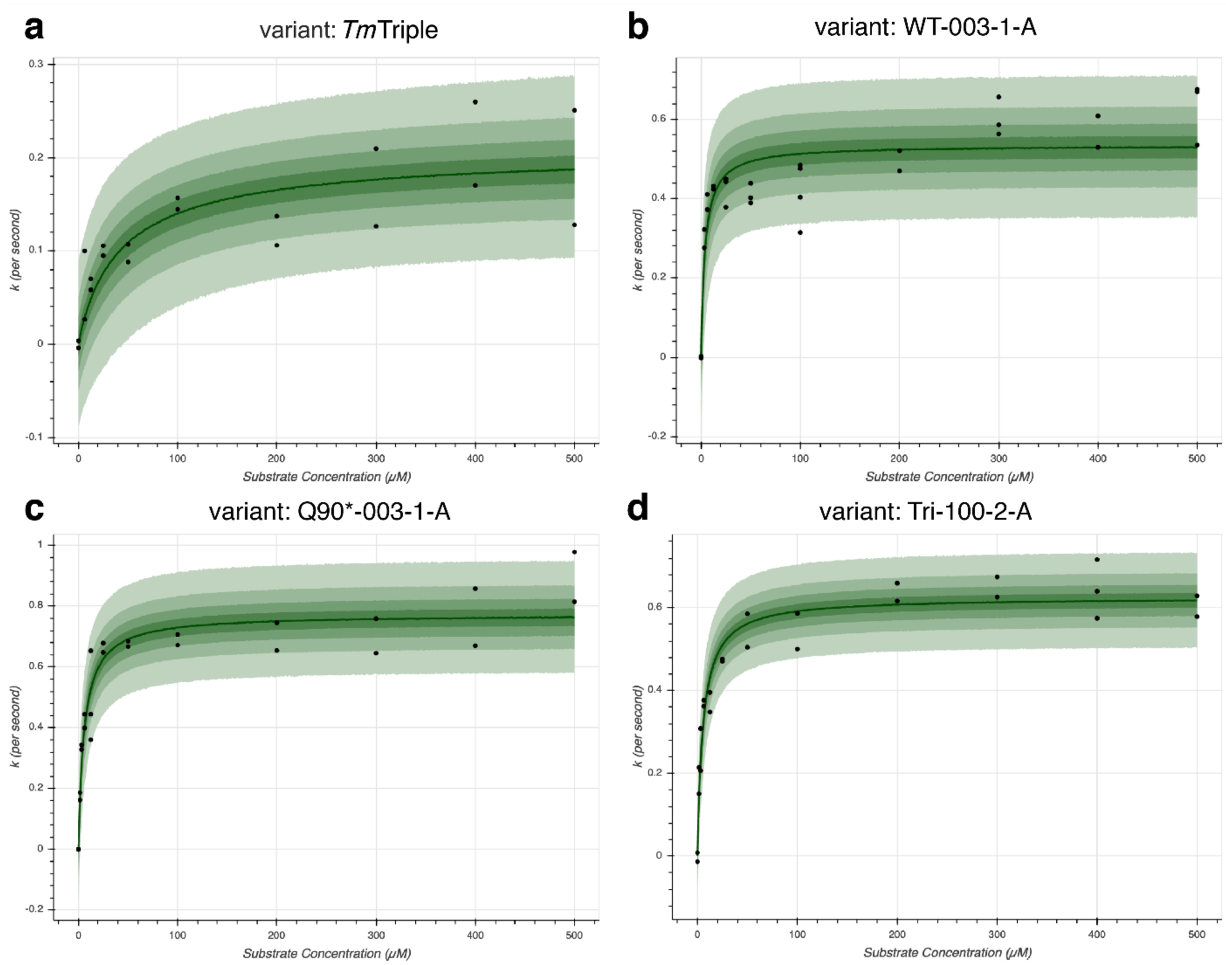


**Figure S2.6 | Michaelis-Menten plots for rate of Trp production at saturating serine for evolved *Tm*TrpB variants.** Initial rate of Trp formation (*k*, per second) with TrpB variants *Tm*Triple (**a**), WT-003-1-A (**b**), Q90*-003-1-A (**c**), or Tri-100-2-A (**d**) at saturating serine concentration (40 mM) vs. indole concentration. Points, median estimates for initial rate based on absorbance change over time (see **Methods**). The median estimated Michaelis-Menten curve is shown as a dark green line, with the 25, 50, 75, and 95% credible regions displayed from dark to light green, respectively. All measurements performed in at least technical duplicate.

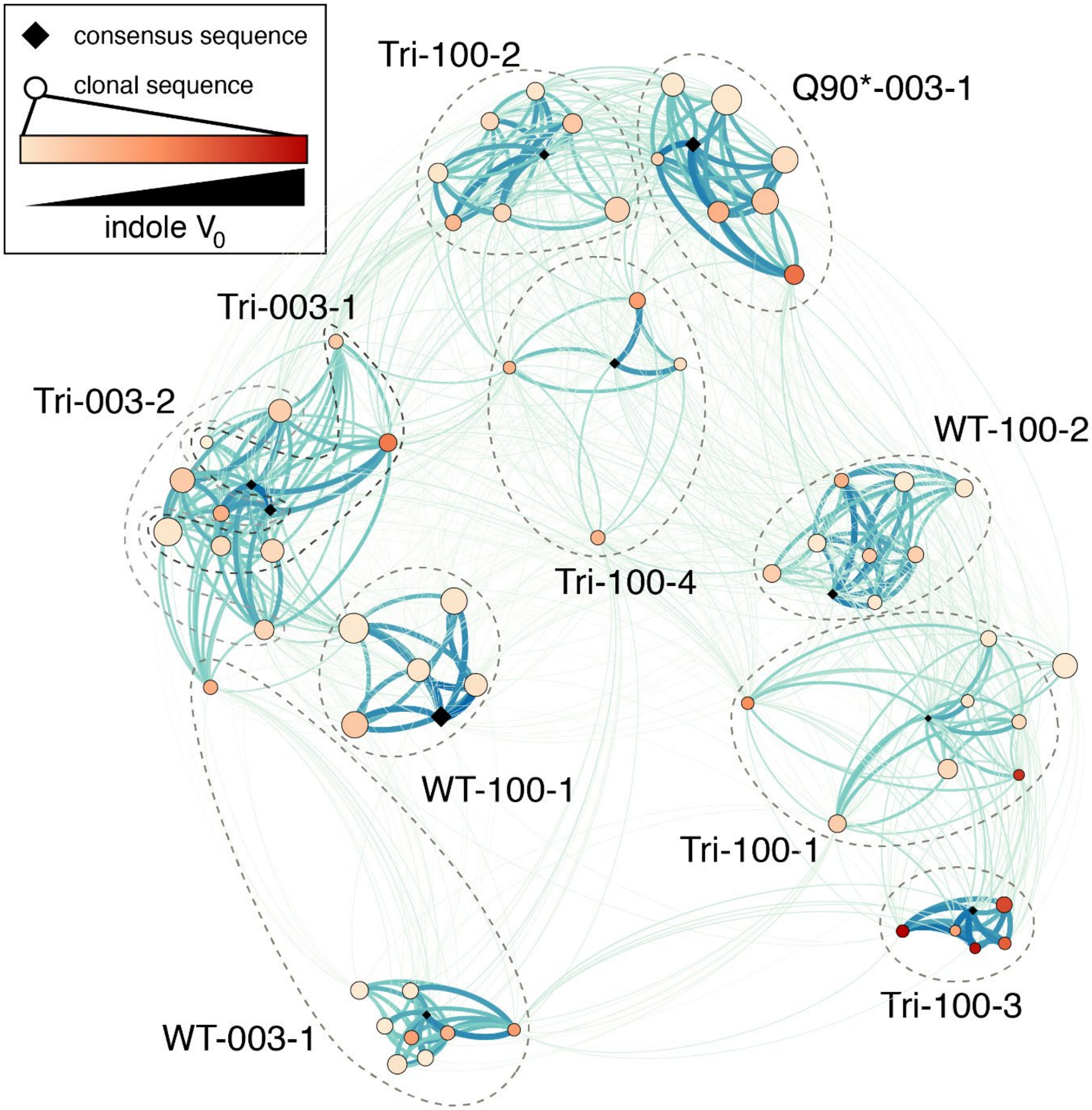


**Figure S2.7 | Relatedness of TrpB panel sequences generated by OrthoRep evolution.** Force directed graph where each node represents an individual sequence (all variants from set 2, and variants WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A) or consensus sequence for one of the ten evolved populations. Edge weights are proportional to the number of shared mutations between two nodes. Higher edge weight yields a stronger attractive force between two nodes, and is visualized as a darker color and a thicker line. Nodes for individual sequences are colored according to initial rate of Trp formation, similar to **Fig. 3b**, and are sized according to the number of mutations in the sequence. Dotted lines are drawn around consensus sequences and individual sequences that are derived from the same evolved culture, if nodes are sufficiently clustered to allow it. Graph was visualized using Gephi version 0.9.2.

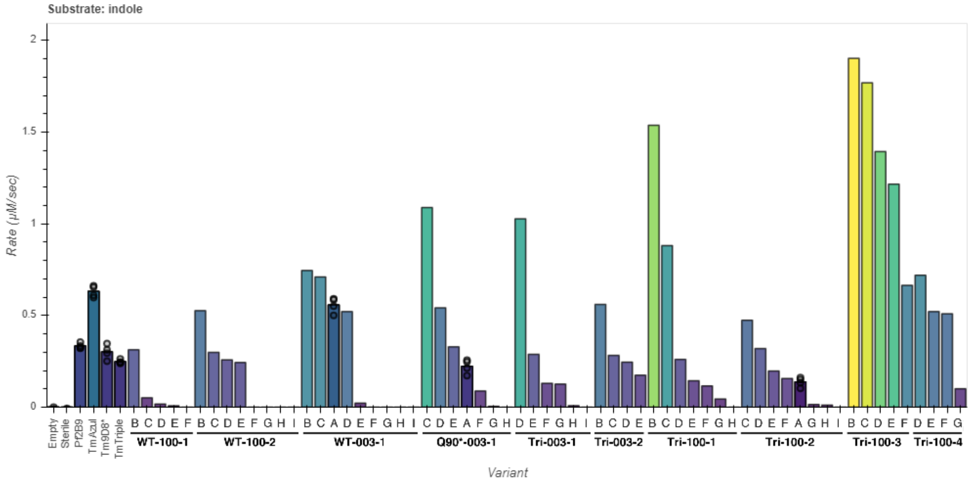
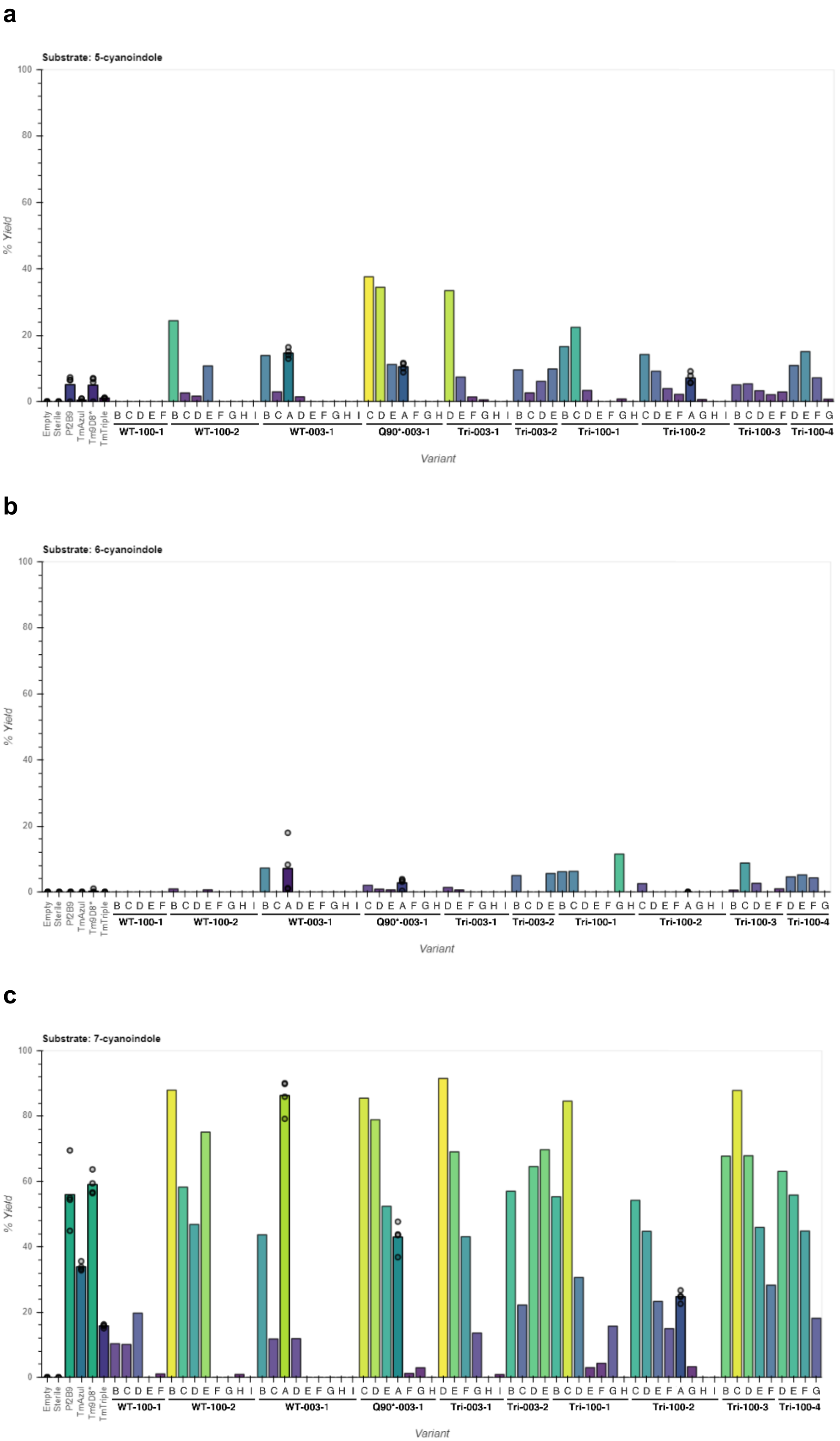
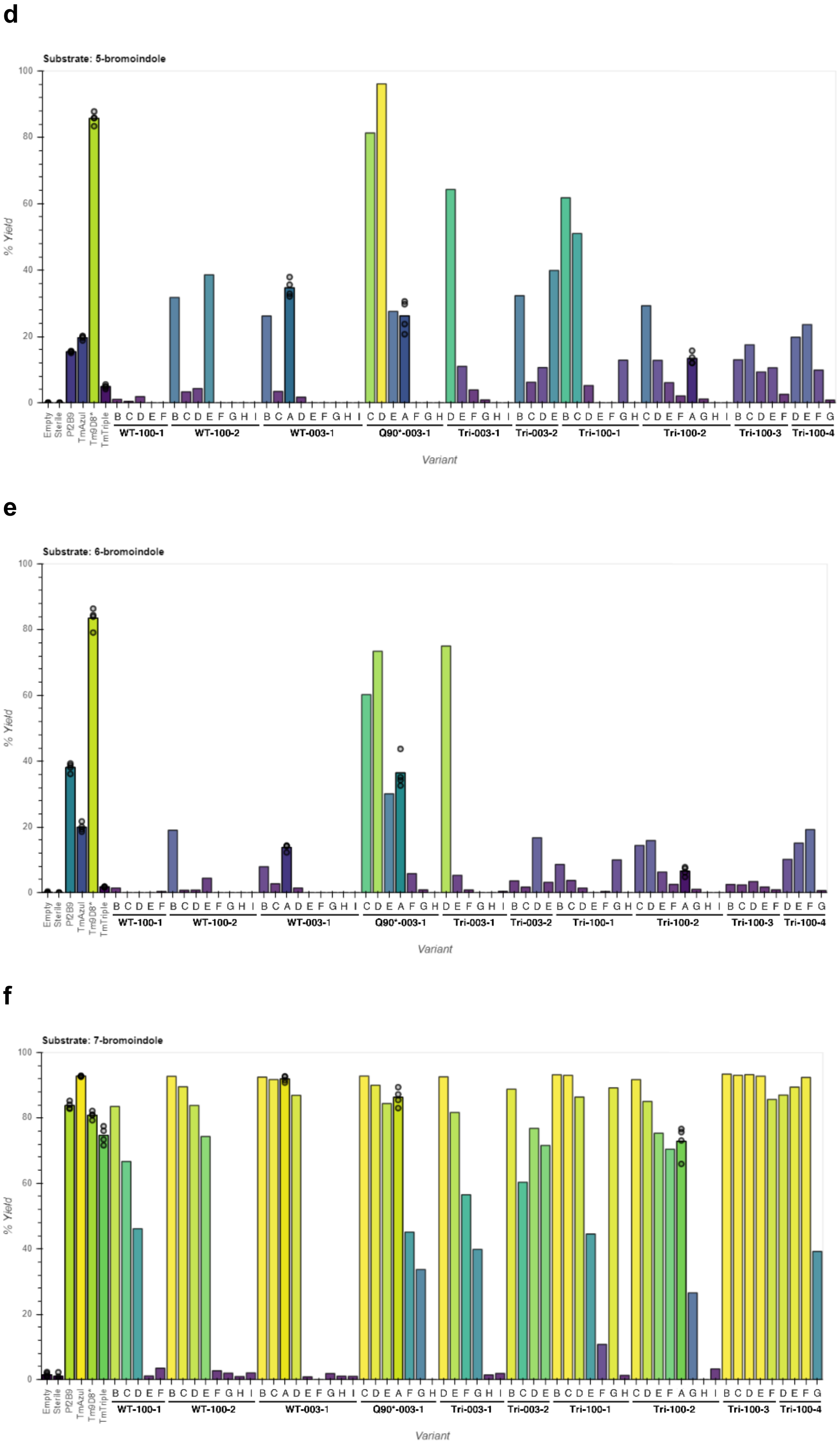
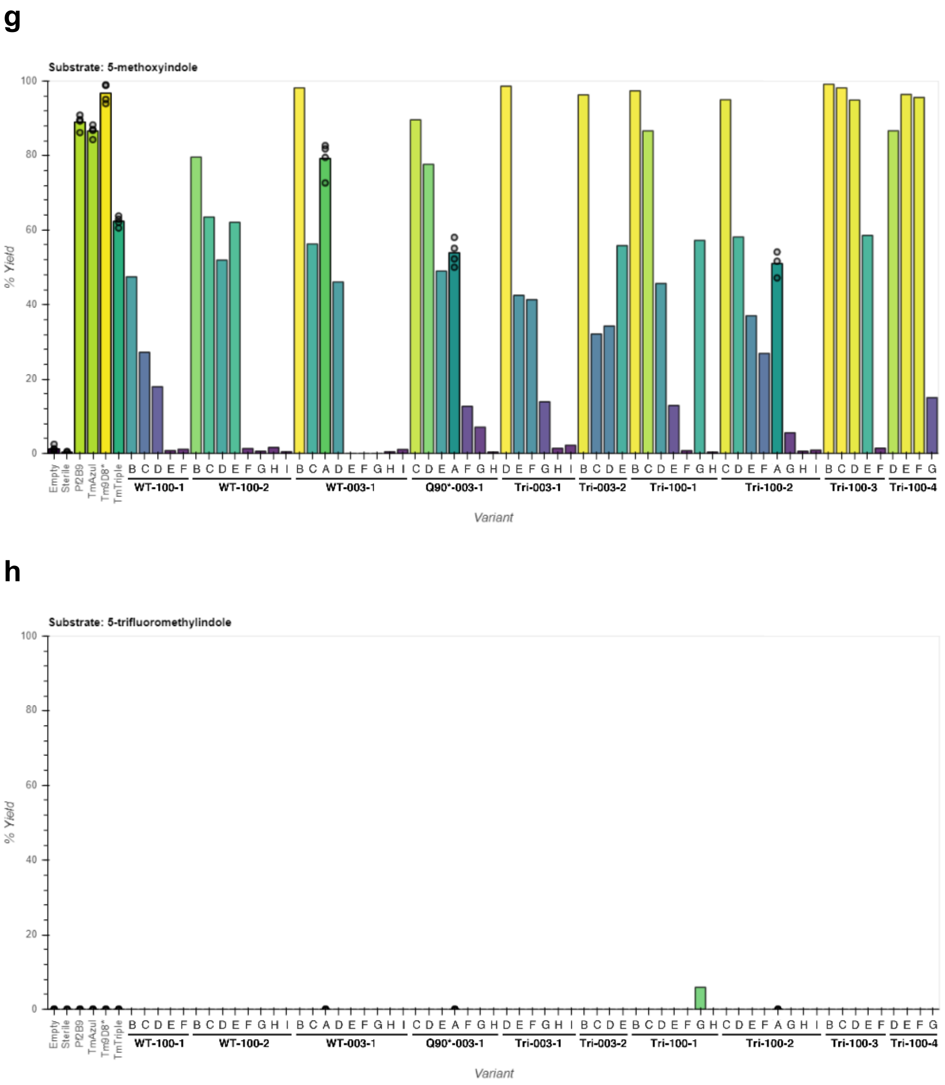


**Figure S2.8 | TrpB panel indole activity by initial rate of Trp formation.** Initial rate of Trp formation at saturating L-serine by UV-vis spectrophotometry. Points represent rate for individual replicates, bars represent mean rate for reactions with multiple replicates, or rate for a single replicate otherwise. OrthoRep-evolved variants are ordered first by the population from which they were derived, then by indole activity. Empty, expression vector without TrpB variant. Sterile, reaction master mix without heat treated lysate added. Empty, *Pf*2B9, *Tm*Azul, *Tm*9D8*, *Tm*Triple, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A all performed in quadruplicate; sterile performed in duplicate; all other reactions performed in a single replicate.


a
Substrate: 5-cyanoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
WT-100-1
WT-100-2
WT-003-1
Q90*-003-1
Tri-003-1
Tri-003-2
Tri-100-1
Tri-100-2
Tri-100-3
Tri-100-4
Variant
b
Substrate: 6-cyanoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
WT-100-1
WT-100-2
WT-003-1
Q90*-003-1
Tri-003-1
Tri-003-2
Tri-100-1
Tri-100-2
Tri-100-3
Tri-100-4
Variant
c
Substrate: 7-cyanoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
WT-100-1
WT-100-2
WT-003-1
Q90*-003-1
Tri-003-1
Tri-003-2
Tri-100-1
Tri-100-2
Tri-100-3
Tri-100-4
Variant


d
Substrate: 5-bromoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
PfZ89
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant
e
Substrate: 6-bromoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
PfZ89
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant
f
Substrate: 7-bromoindole
% Yield
0
20
40
60
80
100
Empty
Sterile
PfZ89
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant


g
Substrate: 5-methoxyindole
% Yield
0
20
40
60
80
100
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant
h
Substrate: 5-trifluoromethylindole
% Yield
0
20
40
60
80
100
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant

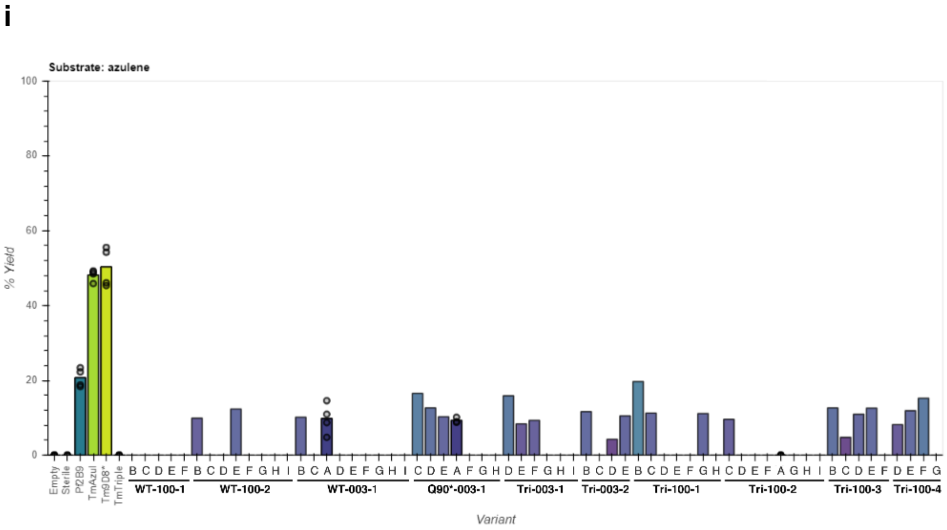


**Figure S2.9 | TrpB panel activity with indole analogs by HPLC yield. a**-**i**, HPLC yield of (**a**) 5-cyanoTrp, (**b**) 6-cyanoTrp, (**c**) 7-cyanoTrp, (**d**) 5-bromoTrp, (**e**) 6-bromoTrp, (**f**) 7-bromoTrp, (**g**) 5-methoxyTrp, (**h**) 5-trifluoromethylTrp, and (**i**) β-(1-azulenyl)-L-alanine (azulene) for indicated variants supplied with L-serine and each indole substrate. Points represent % yield for individual replicates, bars represent mean % yield for reactions with replicates, or % yield for a single replicate otherwise. Replicates and variant order are as in Fig. S8, and populations from which OrthoRep-evolved variants were derived are annotated.

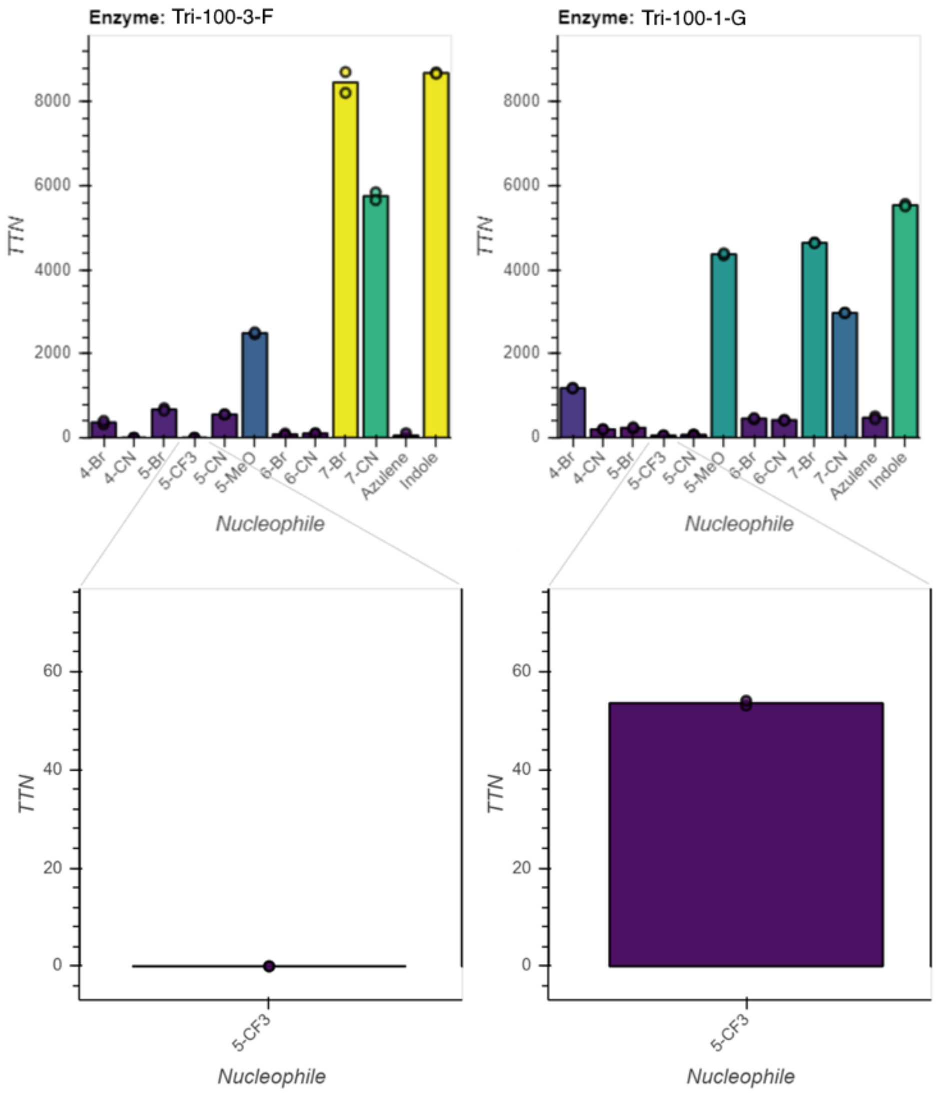


**Figure S2.10 | Substrate activity profiles for large scale purification of variants Tri-100-3-F and Tri-100-1-G.** Total turnover number (TTN) for *Tm*TrpB variants Tri-100-3-F and Tri-100-1-G purified at large scale (see **Methods**) and supplied with L-serine and the indicated indole analog, azulene, or indole (nucleophile), with a maximum TTN of 10,000. All reactions were performed in technical duplicate. Points represent TTN for individual replicates, bars represent mean for two replicates. Insets, TTN for 5-trifluoromethylindole with y-axis scale adjusted for clarity.

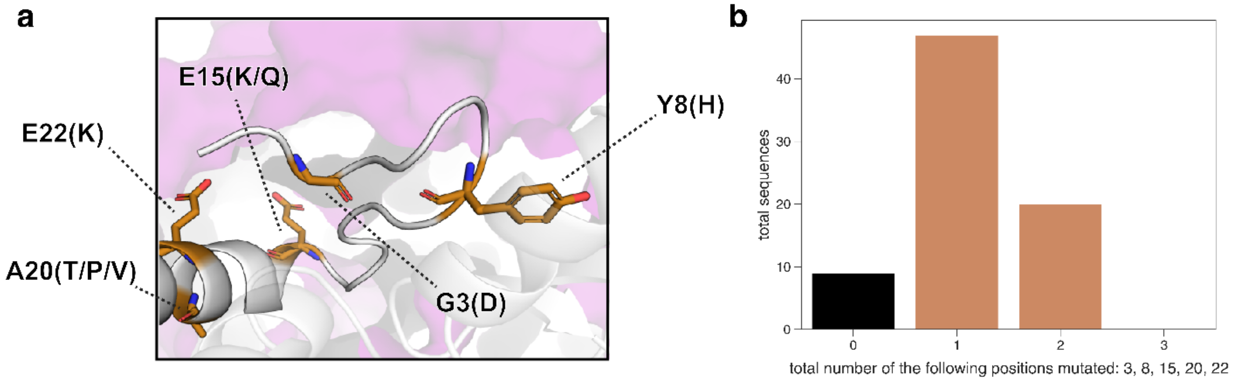


**Figure S2.11 | Commonly observed mutations at the α-subunit interaction interface. a**, Homology-predicted *Tm*TrpB structure (based on engineered stand-alone *Pf*TrpB, PDB 6AM8), with commonly mutated *Tm*TrpB residues located near the TrpA interaction interface highlighted. Solvent-exposed regions of TrpA (purple) (PDB 1WDW) are shown as a surface. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. **b**, Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panel **a**.

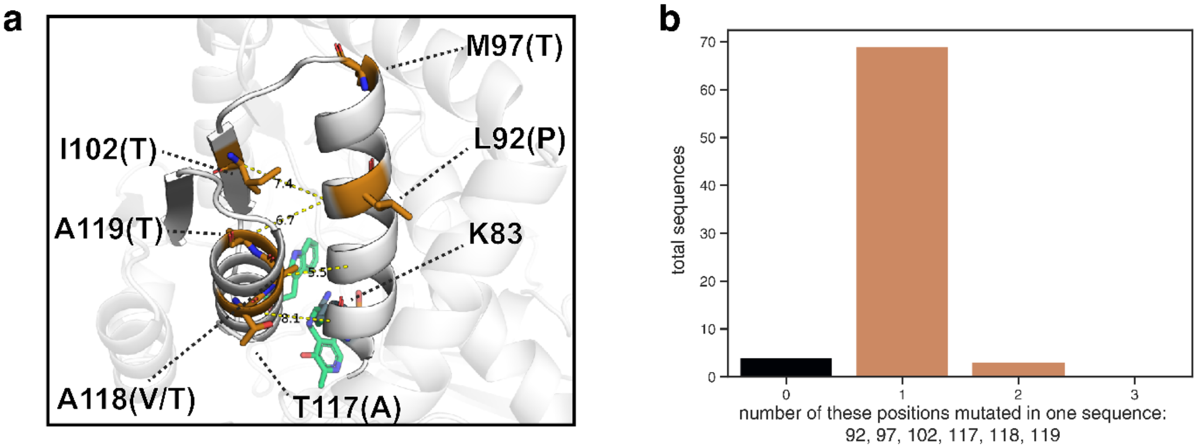


**Figure S2.12 | Commonly observed mutations to residues near a catalytic α-helix. a**, Homology-predicted *Tm*TrpB structure (aligned to engineered stand-alone *Pf*TrpB, PDB 6AM8) with wt residues on or near the α-helix housing K83, which are commonly mutated in OrthoRep evolved populations (orange). PLP (green) and Trp (green) are shown as sticks, and the catalytic lysine K83 (teal) is shown as spheres. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. Dotted lines connect the α-carbon of residues not located on the K83 α-helix with the α-carbon of the nearest residue on the K83 α-helix, with the distance noted in Ångstroms. **b**, Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panel **a**.

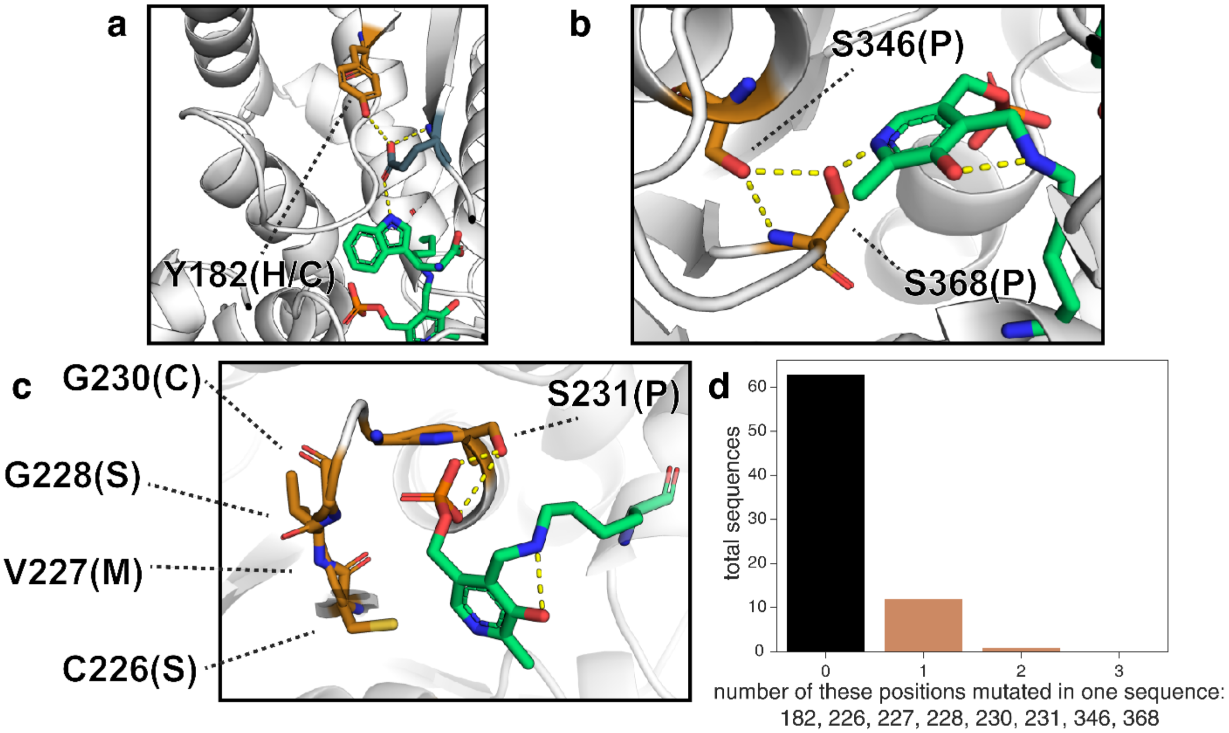


**Figure S2.13 | First- and second-shell active site mutations. a-c**, Homology model of *Tm*TrpB (aligned to *Pf*TrpB, PDB: 6AM8) highlighting residues mutated in OrthoRep-evolved variants (orange) that may influence (**a**) indole charge, (**b**) PLP six-member ring binding, and (**c**) PLP-phosphate binding. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. **d**, Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panels **a**, **b**, or **c**.

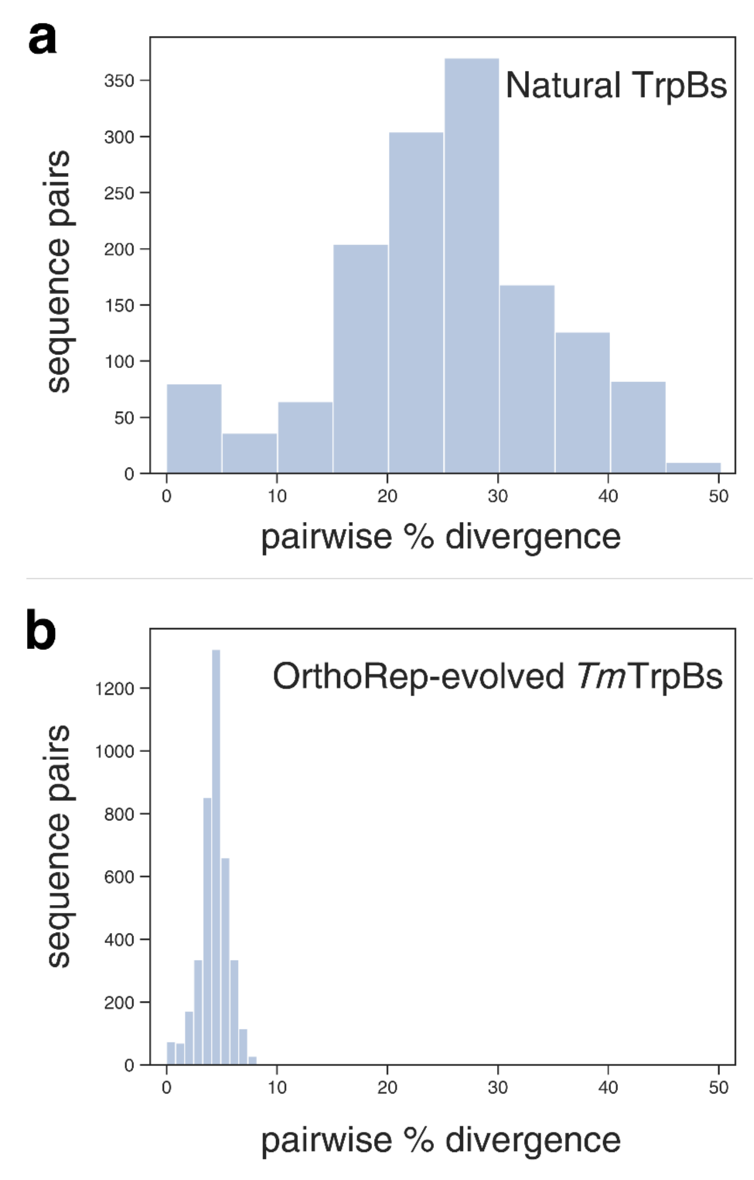


**Figure S2.14 | Sequence divergence for natural and OrthoRep-evolved TrpBs. a-b**, Distributions of pairwise % amino acid sequence divergence for a diverse group of 38 naturally occurring mesophilic TrpB variants (**a**) and OrthoRep-evolved variant sets 1 and 2 (**b**).

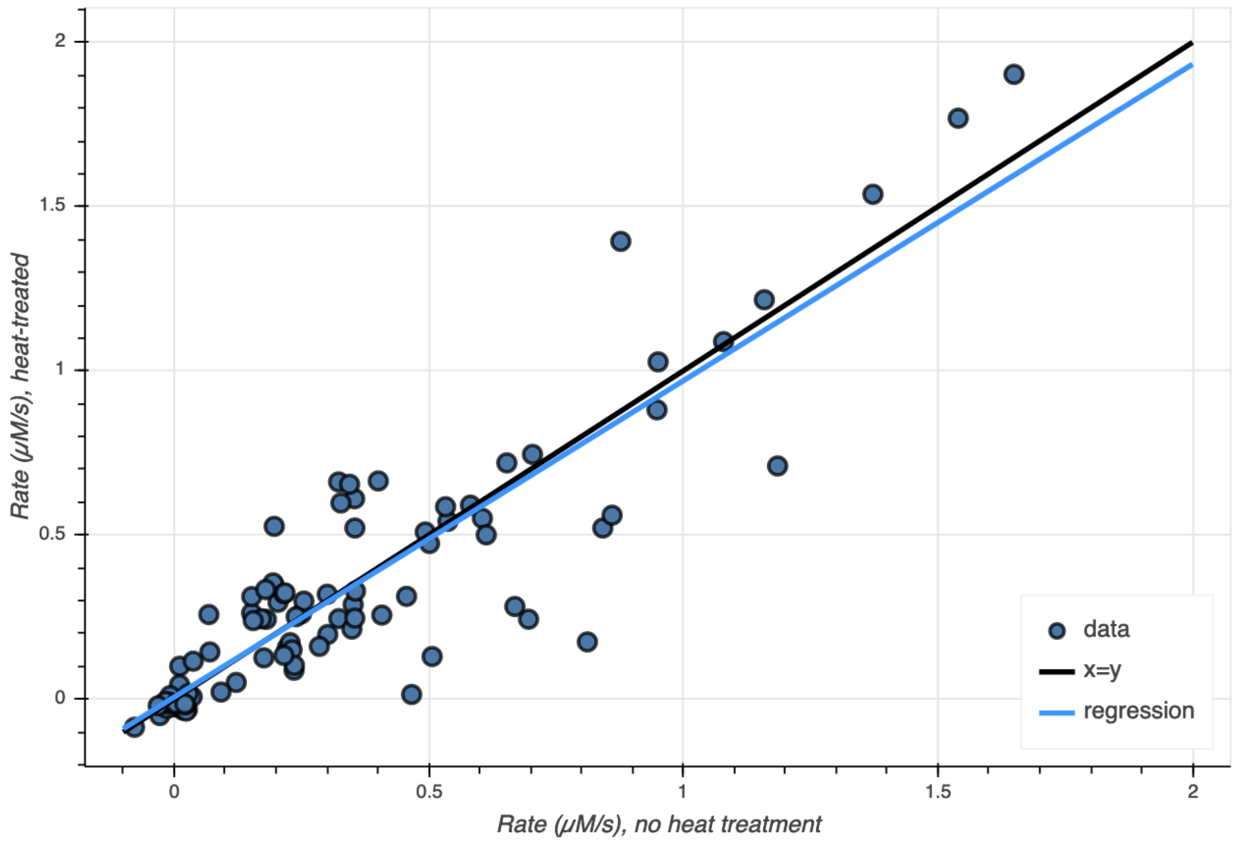


**Figure S2.15 | Correlation between Trp formation by lysates with and without heat treatment.** Trp formation by each OrthoRep-evolved variant from variant set 2, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A with or without heat treatment at 625 µM indole and 25 mM serine, evaluated by UV-vis spectrophotometry. Each point represents the initial rate of Trp formation by cell lysate generated via both heat treatment (1 hour at 75 °C) (y-axis) and a more mild method (x-axis) (see **Methods**). Linear regression on these data demonstrates a slope of 0.96, suggesting a negligible systematic decrease in activity with heat treatment across variants.

# Appendix B: Supplementary information for Chapter 3

## Supplementary Figures

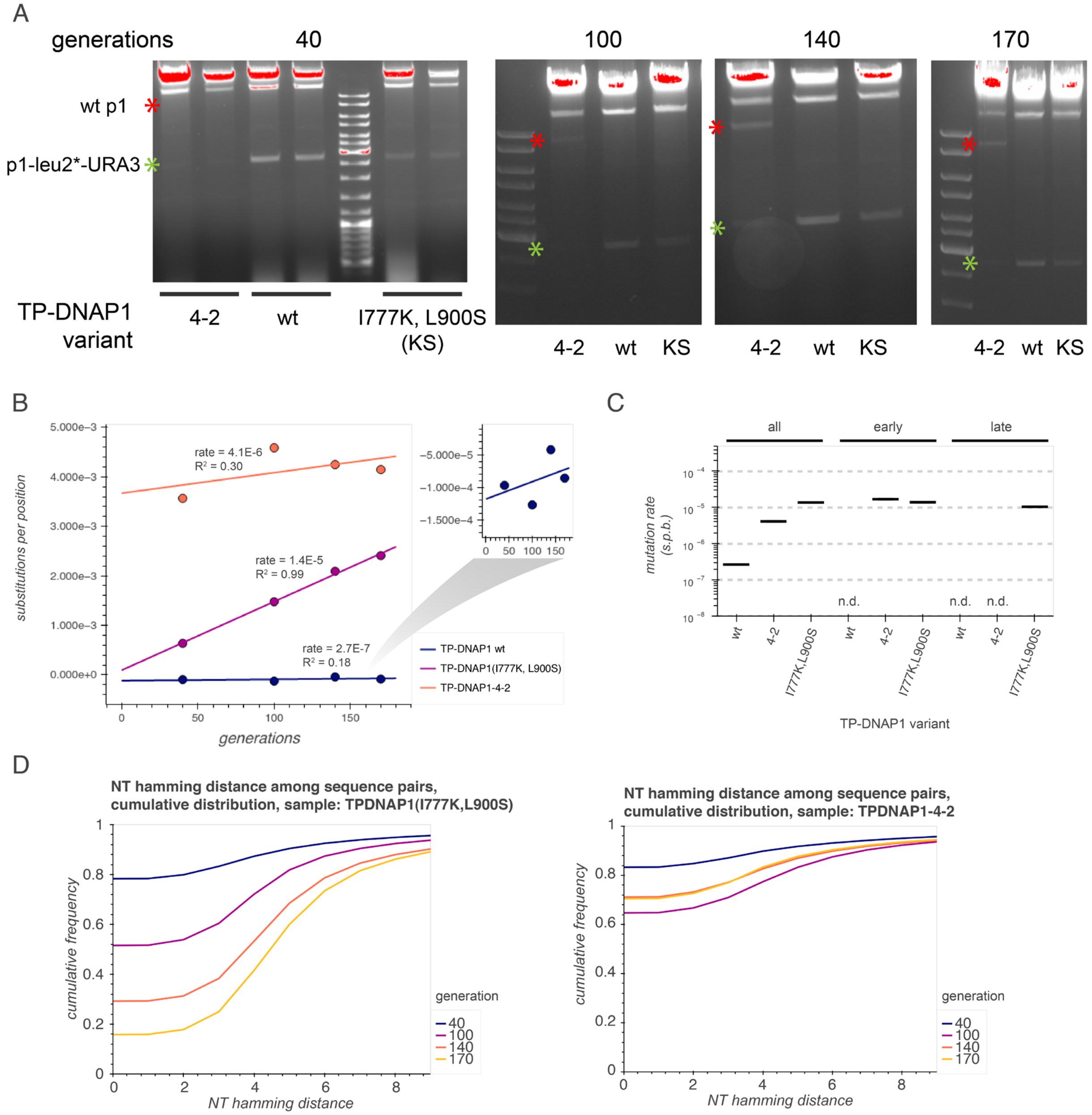


**Fig. S1. Validation of mutation accumulation and comparison of p1 maintenance by legacy TP-DNAP1 variants.** (**A-D**) Yeast strains encoding p1-leu2*-URA3 were transformed with plasmids encoding wild-type (wt) TP-DNAP1, TP-DNAP1-4-2, or TP-DNAP1(I777K, L900S)

and passaged for 130 generations under selection for URA3 to allow for accumulation of mutations in leu2*. DNA was isolated from these samples at four timepoints throughout the experiment. Gel electrophoresis of these DNA samples (A) revealed the resurgence of a DNA band corresponding in length to wt p1 (~9 kb) in the TP-DNAP1-4-2 sample, but not others. Timepoint DNA isolates were used for PCR amplification, high throughput sequencing, and mutation analysis of a ~450 bp region demonstrated that TP-DNAP1(I777K, L900S) maintained both a consistent mutation rate (B, C) and monotonic diversification throughout the experiment (D). Two replicates were isolated for the first timepoint for gel electrophoresis, but only one of these was isolated for subsequent timepoints and used for mutation accumulation. Mutations per base for WT is shown rescaled in (B) to highlight the poor linear fit. n.d., not detected.

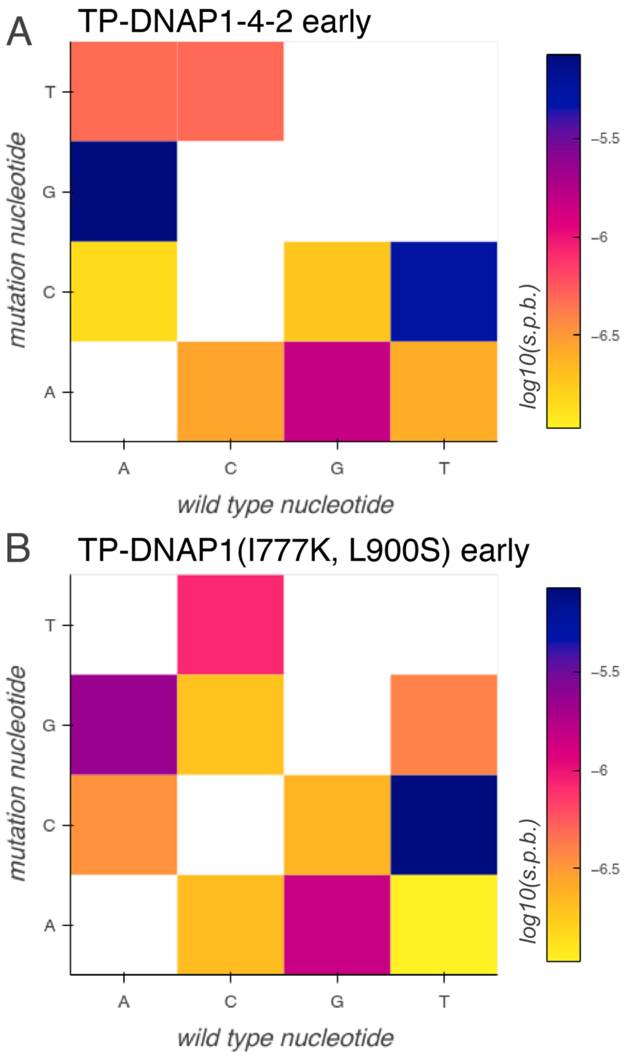


**Fig. S2. Mutation spectrum of legacy TP-DNAP1s.** (**A-B**) Heatmaps of log transformed individual substitution rates for all 12 substitution types for TP-DNAP1-4-2 (A) and TP-DNAP1(I777K, L900S) (B) from the first two timepoints of mutation accumulation in HTS dataset I (~60 generations, see Fig. S1).

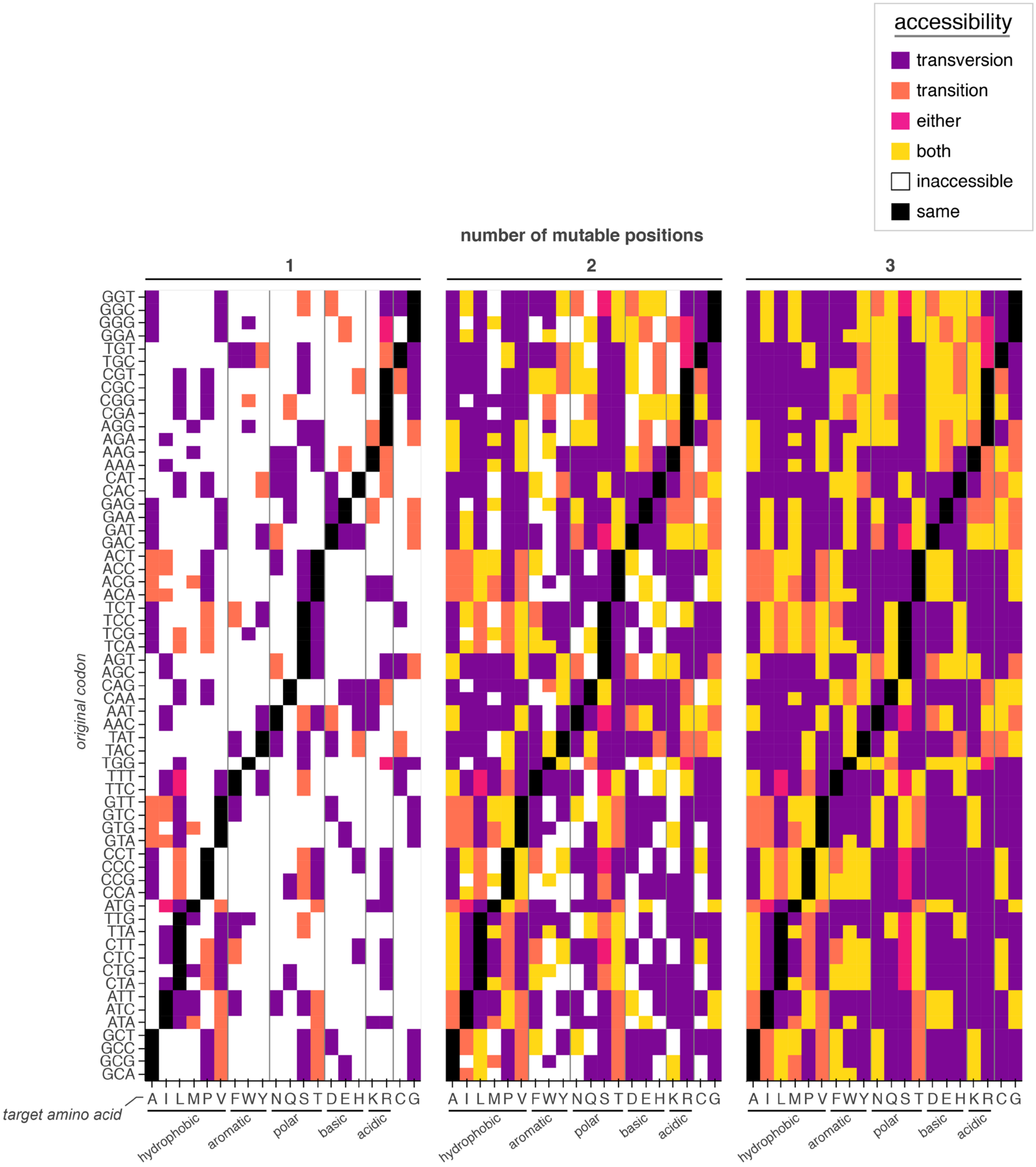


**Fig. S3. Amino acid accessibility by mutation type.** Accessibility of any codon for each amino acid from each of the 64 possible codons by one (left column), two (middle column), or three (right column) nucleotide mutations. Color indicates whether the codon can reach the corresponding amino acid with only transversions (purple), with only transitions (salmon), either with only transversions or with only transitions (pink), only with both transversions and transitions (yellow), or is inaccessible (white) with at most the indicated number of mutable positions. Multiple mutations to the same position within the codon were not considered.

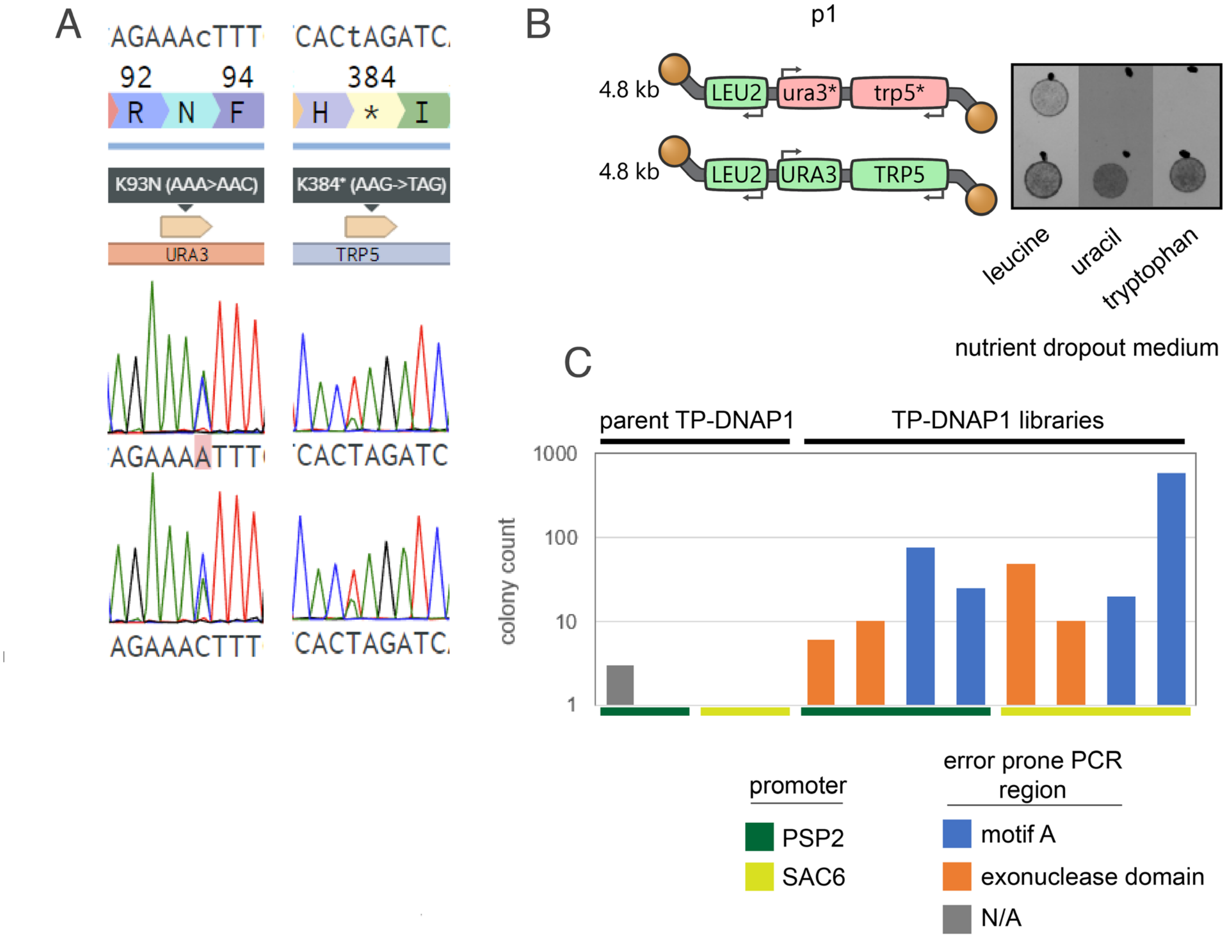


**Fig. S4. Validation of transversion-specific selection and epPCR 1.** (**A**) Sanger sequencing of the ura3-K93N (ura3*) or trp5-K384* (trp5*) transversion mutations in two distinct clonal populations of yeast harboring the p1-ura3*-trp5* plasmid following selection in -uracil or -tryptophan media, respectively. In all cases, the only visible peaks were those corresponding to the original sequence or the desired transversion mutation. 100% fixation of the expected mutation was not enforced due to the multicopy nature of p1. (**B**) Plating assay demonstrating that p1-encoded auxotrophy markers URA3 and TRP5 were rendered inactive by the K93N and K384* mutations to catalytic lysines. (**C**) URA3 revertant colony counts for either TP-DNAP1 epPCR 1 libraries or the parent TP-DNAP1(I777K, L900S) transformed into OR-Y488. TP-DNAP1(I777K, L900S) was used as the template sequence for error prone PCR mutagenesis of one of two regions of the polymerase expressed under one of two promoters for a total of four $10^3$ – $10^4$ member TP-DNAP1 libraries which were then transformed into the OR-Y488 strainharboring p1-ura3*-trp5*. Two biological replicates of each of these libraries were subject to selection for URA3 alongside mock libraries encoding only the parental TP-DNAP1 sequence and resulting revertants were quantified by titering and replating.

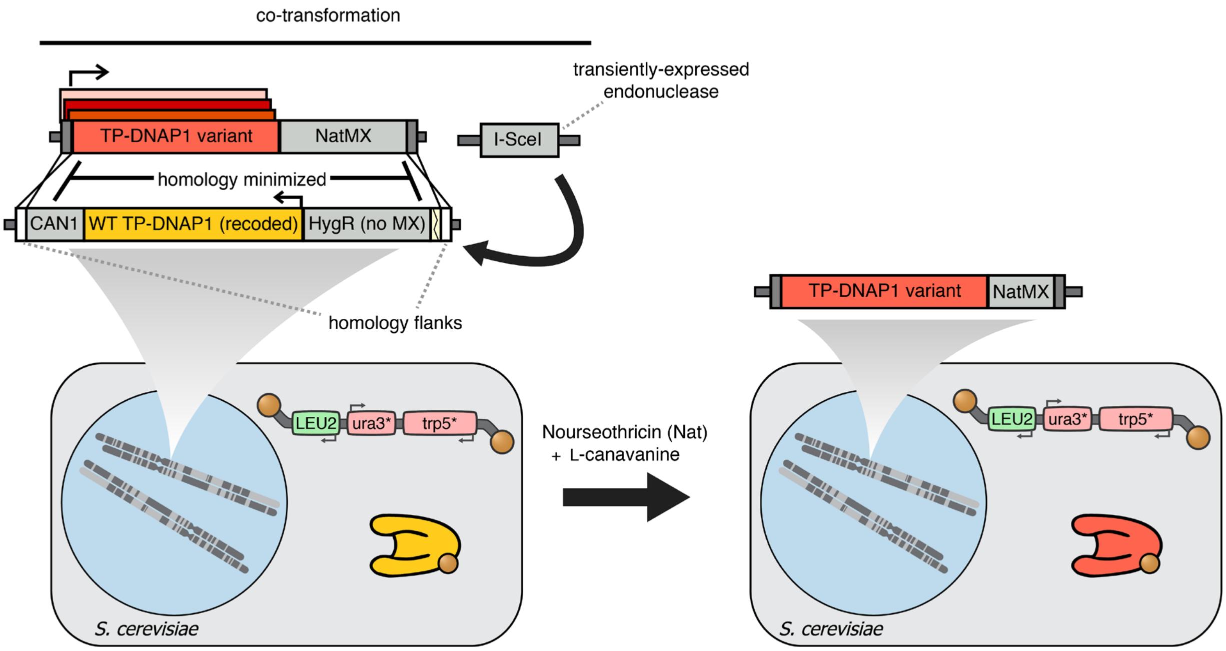


**Fig. S5. Illustration of the design of the polymerase replacement transformation.** A strain encoding a "landing pad" DNA sequence that includes the wild type (WT) TP-DNAP1 and an I-SceI cut site at the CAN1 locus was co-transformed with both a TP-DNAP1 variant (typically in library format) and a transient I-SceI endonuclease expression cassette[15]. Integration of NatMX was selected for using nourseothricin. Retention of the CAN1 in the landing pad was selected against using L-canavanine.

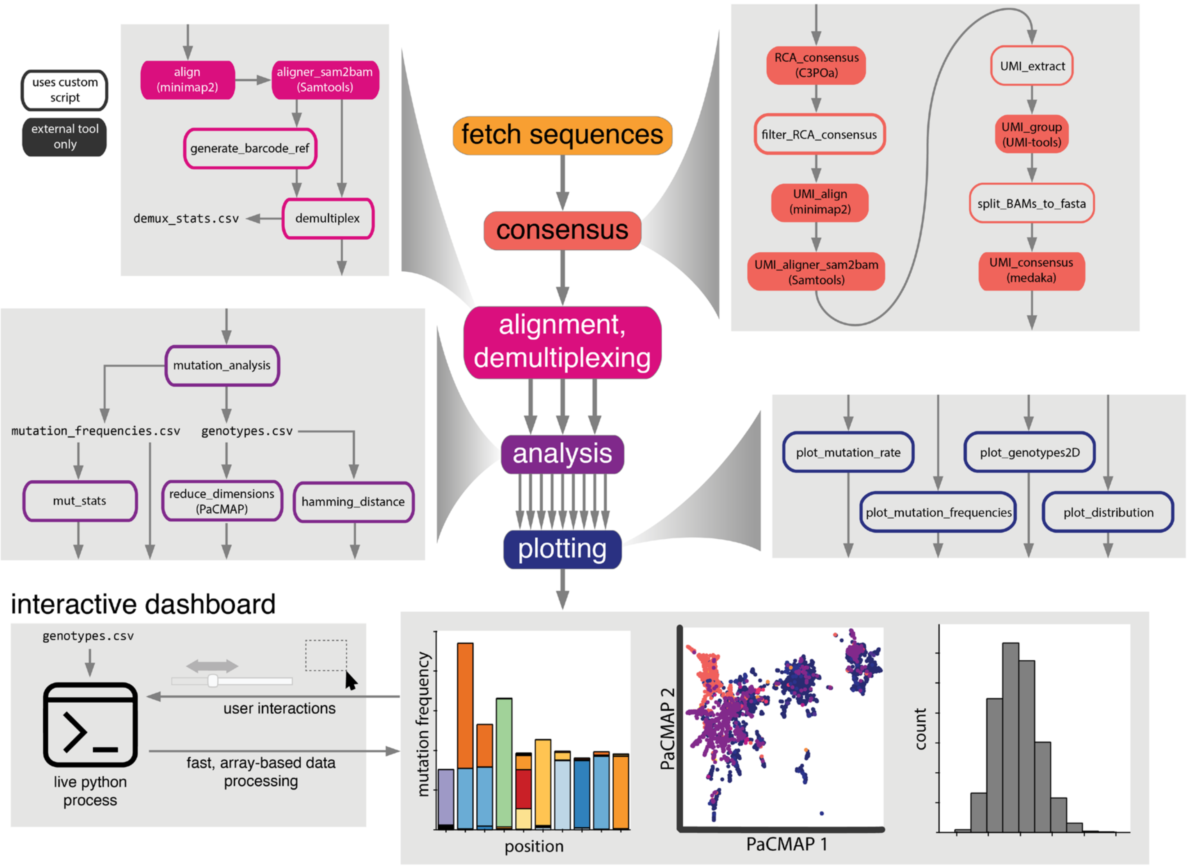


**Fig. S6. Flowchart of the programmatic steps performed by the mutation analysis for parallel laboratory evolution (Maple) pipeline.** To facilitate rapid exploration and analysis of high throughput sequencing datasets, we built an end-to-end data processing pipeline that takes as input a sequencing dataset and a minimal set of additional user inputs and carries out the necessary steps to produce commonly desired visualizations as well as the data that supports those visualizations. This includes generating high accuracy consensus sequences from multiple reads of a sequence via rolling circle amplification (RCA) or unique molecular identifiers (UMI), alignment-based demultiplexing to separate and label sequences derived from different samples, and mutation analysis to generate human-readable .csv outputs that are further analyzed and visualized by Maple or can be viewed and analyzed by the user. Maple also includes an interactive dashboard that allows for user interaction with visualizations, such as selection of sequences that cluster together when plotted by the output of the dimensionality reduction tool PaCMAP.

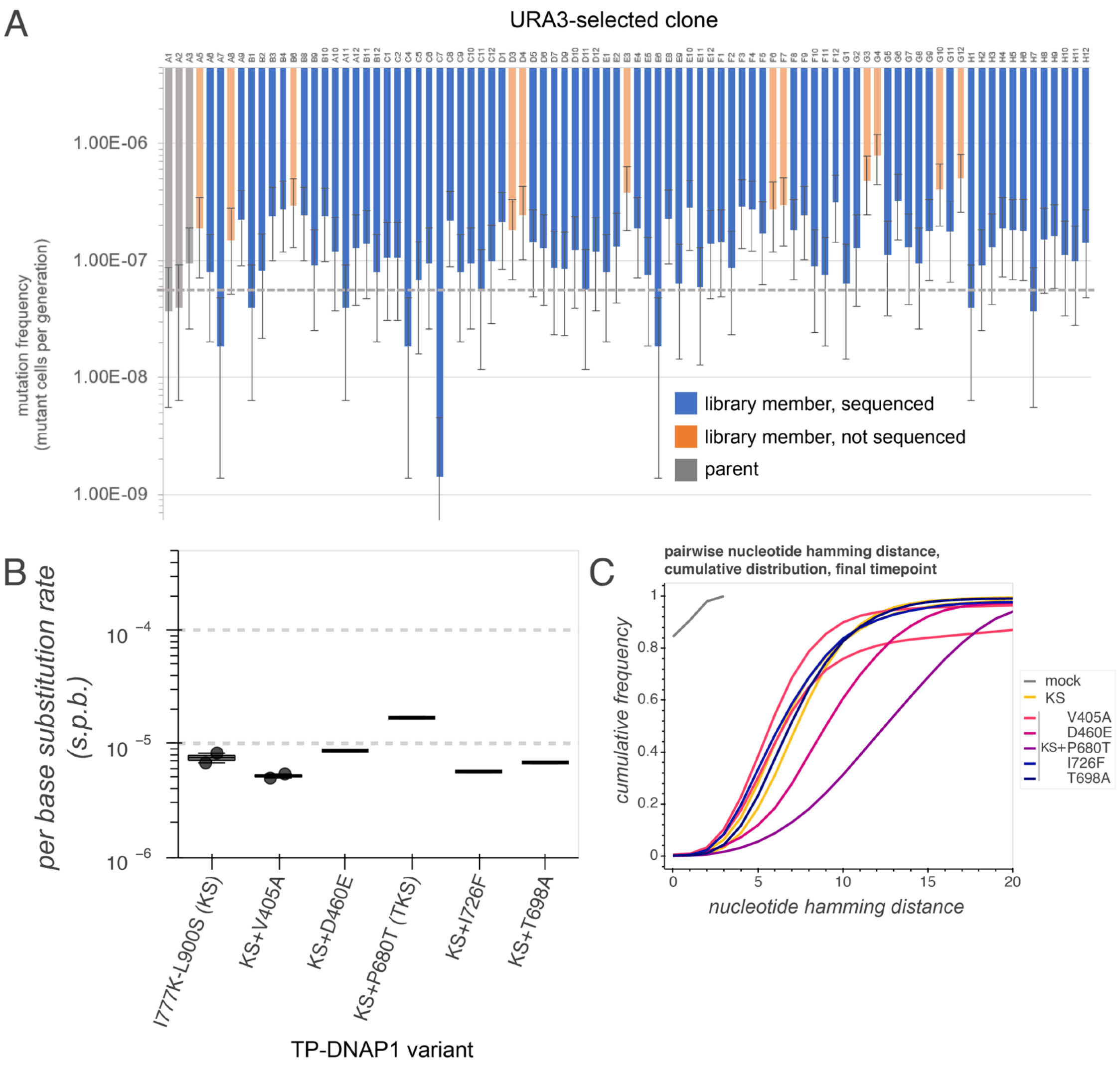


**Fig. S7. Identification of TP-DNAP1-TKS (epPCR 1).** (**A**) Fluctuation assay results for ~90 colonies isolated from OR-Y488 transformed with either the epPCR 1 library (blue/orange) or the parent TP-DNAP1(I777K, L900S) (grey), then subject to URA3 selection. Note that per base mutation rate cannot be calculated by fluctuation analysis without copy number measurement; thus only per-cell mutation frequency is shown. Twelve library members (orange) were chosen for Sanger sequencing, using both mutation frequency and lineage (to minimize duplicate TP-DNAP1 variants) as criteria. Grey dotted line, mean mutation frequency for the three parent controls. (**B**) Mutation rate measurements for TP-DNAP1(I777K, L900S) and five unique TP-DNAP1 variants isolated from this directed evolution round. Following selection for URA3 and either fluctuation analysis (using trp5) or trp5 selection, individual clones were grown in -uracil growth medium for mutation accumulation on trp5. Two timepoints separated by 100 generations of growth were used as template for long-read high-throughput sequencing. Mutation rates are shown as a box plot and points for individual replicates where n > 1 biological replicate, or one line where n = 1. (**C**) Cumulative hamming distance distribution for all high throughput sequencing samples at the end of mutation accumulation.

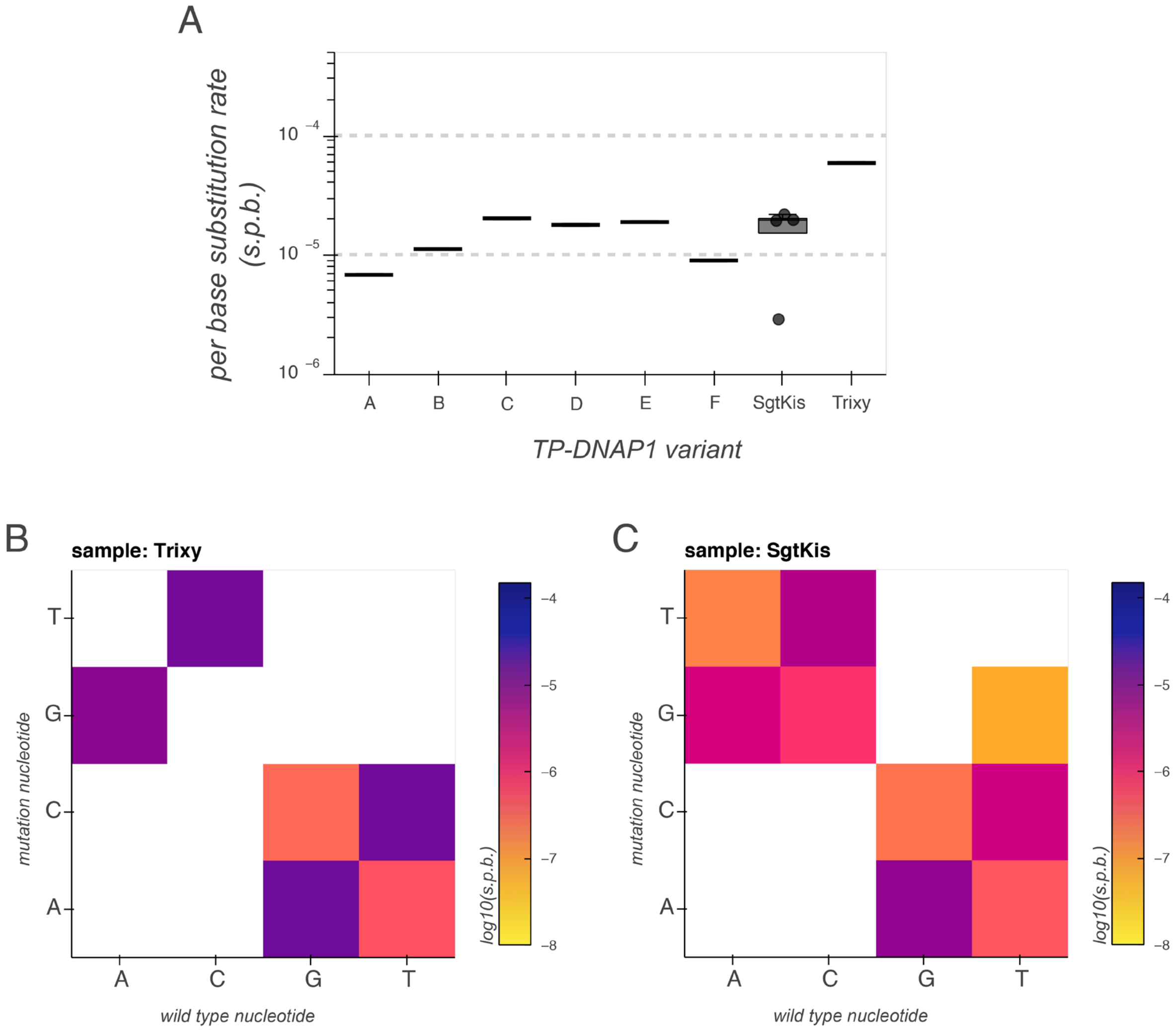


**Fig. S8. Identification of Trixy and SgtKis (epPCR 2).** (**A**) Mutation rate measurements for all substitutions from HTS dataset 3 by mutation accumulation. Mutation rates are shown as a box plot and points for individual replicates where n > 1 biological replicate, or one line where n = 1. (**B-C**) Heatmap representation of individual mutation rates for Trixy (B) and SgtKis (C) as measured in HTS dataset 3.

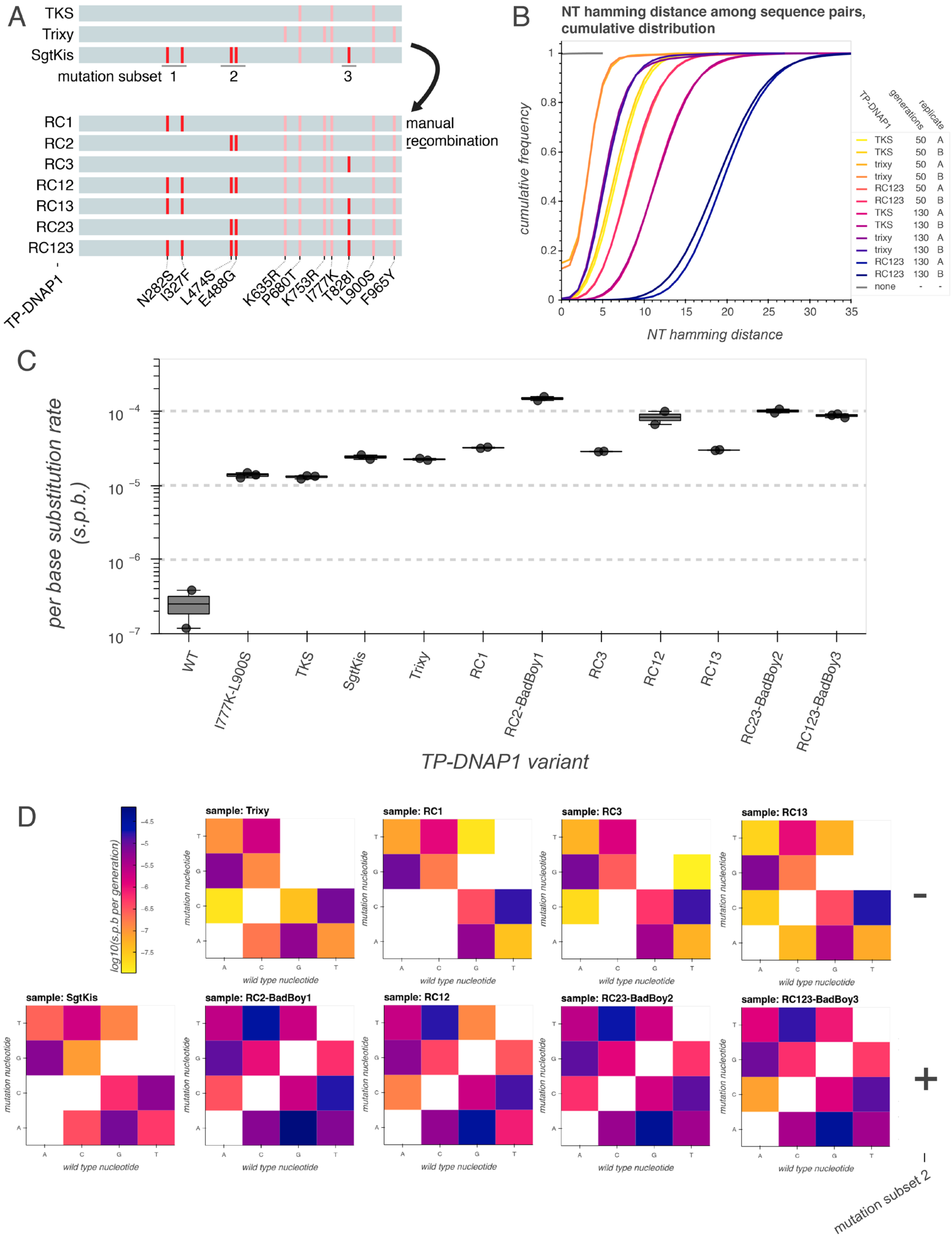


**Fig. S9. Mutation accumulation and HTS analysis of TP-DNAP1 variants from manual recombination.** (**A**) Illustration of manually constructed TP-DNAP1 variants derived from

mutation subsets of SgtKis transplanted onto Trixy. (**B-D**) HTS analysis from an 80 generation mutation accumulation on a ~1 kb region of p1, highlighting the cumulative pairwise hamming distance for a subset of polymerase variants tested (B), the overall mutation rate measurements for all polymerases tested (C), and the effect of mutation subset 2 on the mutation spectrum by log transformed heatmap of individual mutation rates (D). Mutation rates in (B) are plotted as points for rates measured for each replicate and box plots of summary statistics for all biological replicates (n ≥ 2).

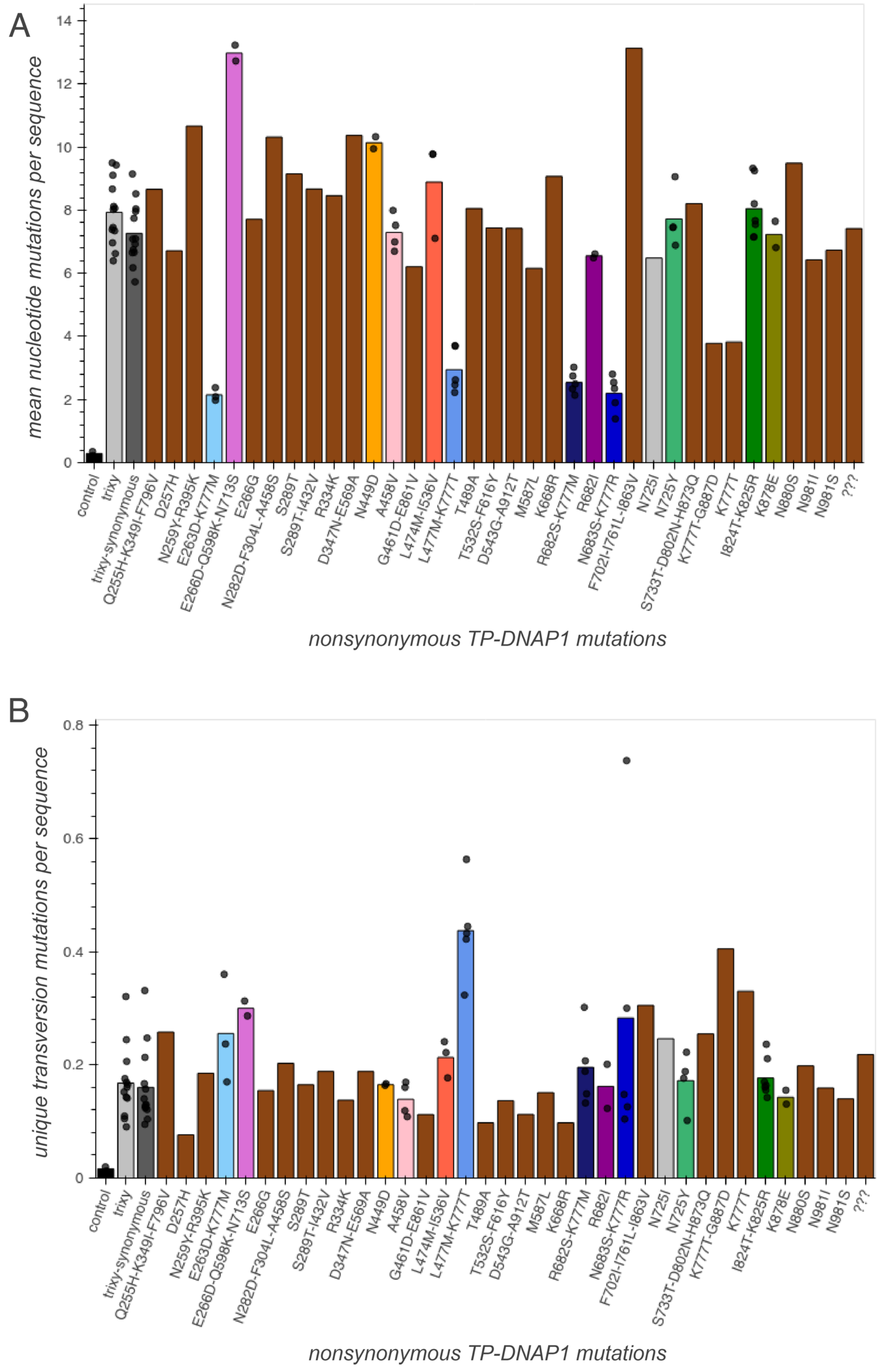

A
mean nucleotide mutations per sequence
0
2
4
6
8
10
12
14
nonsynonymous TP-DNAP1 mutations
B
unique transversion mutations per sequence
0
0.2
0.4
0.6
0.8
nonsynonymous TP-DNAP1 mutations
control
trixy
trixy-synonymous
Q255H-K349I-F796V
D257H
N259Y-R395K
E263D-K777M
E266D-Q598K-N713S
E266G
N282D-F304L-A458S
S289T
S289T-I432V
R334K
D347N-E569A
N449D
A458V
G461D-E861V
L474M-I536V
L477M-K777T
T489A
T532S-F616Y
D543G-A912T
M587L
K668R
R682S-K777M
R682I
N683S-K777R
F702I-I761L-I863V
N725I
N725Y
S733T-D802N-H873Q
K777T-G887D
K777T
I824T-K825R
K878E
N880S
N981I
N981S
???

**Fig. S10. Single timepoint HTS analysis of epPCR 3.** (**A-B**) An error prone PCR library (TP-DNAP1-Trixy template) was subject to either simultaneous or sequential selection for ura3* and trp5* reversion and a ~2kb region of p1 was PCR amplified and subject to HTS. Analysis of mean nucleotide (NT) mutations per sequence (A) and unique transversions per sequence (total unique transversion mutations / number of sequences analyzed) (B) was used to nominate mutations for the following round of directed evolution. Nonsynonymous amino acid substitutions in addition to those found in trixy are listed. Grey, no nonsynonymous TP-DNAP1 mutations in addition to tTP-DNAP1-Trixy mutations identified. Colors, multiple isolates contained the same polymerase. Brown, isolate contained a unique mutation combination. Blues, contains a mutation to residue 777.

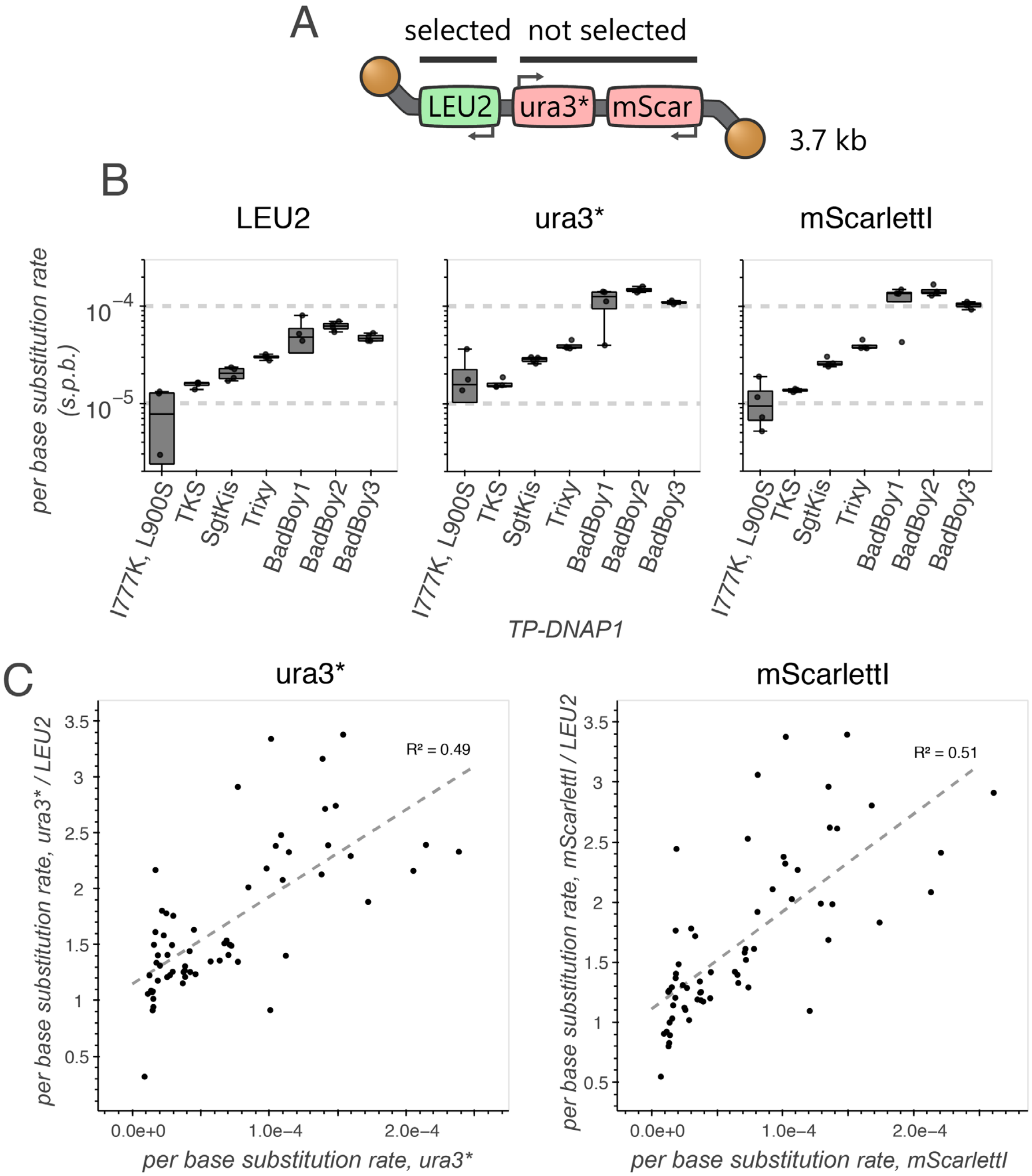


**Fig. S11. The effect of selection on mutation rate.** (**A**) Plasmid map for the p1 subjected to mutation accumulation in HTS dataset 6. Cells were passaged in +uracil/-leucine synthetic complete media resulting in functional selection for LEU2, but not ura3. (**B-C**) Comparison of mutation rates (HTS dataset 6) for regions of similar length (~1 kb), either under selection (LEU2) or not under selection (ura3*, mScarlettI) showing substitution rates for all substitution types as box plots and points (n=4 biological replicates) for a subset of TP-DNAP1 variants (B) or as a scatter plot (C) showing the relationship between mutation rate and fold change in mutation rate without (ura3* or mScarlettI) vs. with (LEU2) selection (one point per each n=1 individual polymerase variant replicate for all 17 TP-DNAP1 variants assayed).

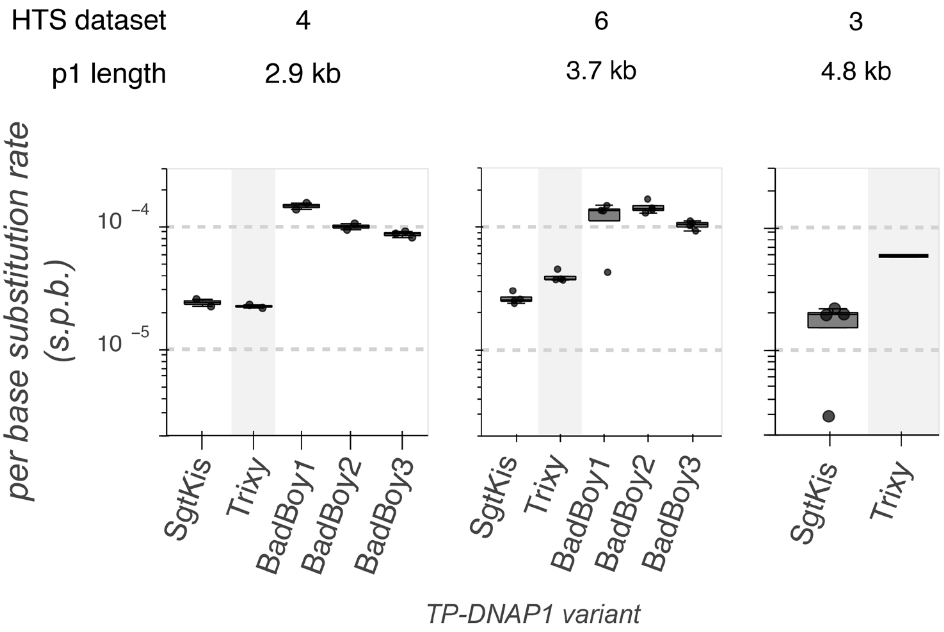


**Fig. S12. Relationship between p1 length and mutation rate.** Mutation rate measurements for a subset of TP-DNAP1 variants and the length of recombinant p1 used to generate the indicated mutation accumulation dataset are shown. TP-DNAP1-Trixy, whose mutation rates show the most obvious relationship with p1 length, is highlighted in grey. Data are identical to those found in Figs. S9, 1C, and S8 (left to right).

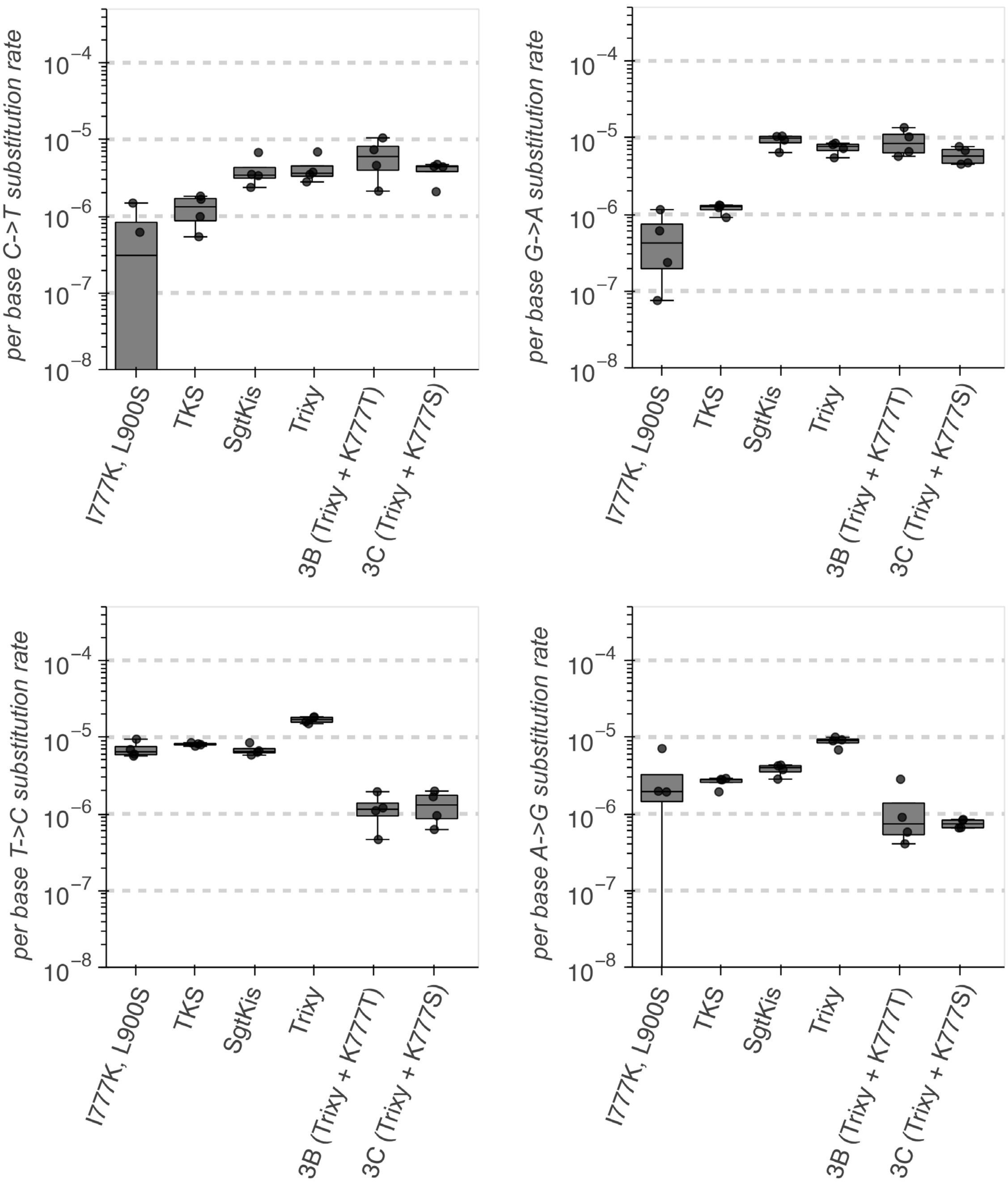


**Fig. S13. The effect of mutations to residue 777 on transition mutation rates.** Mutation rate measurements from HTS dataset 6 for all TP-DNAP1 variants assayed for all four transition mutations, highlighting variants 3B, 3C, which differ from Trixy only by a single mutation (K777T/K777S, respectively) and BB-3B and BB-3C, which differ from BadBoy3 by only a single mutation (K777T/K777S, respectively).

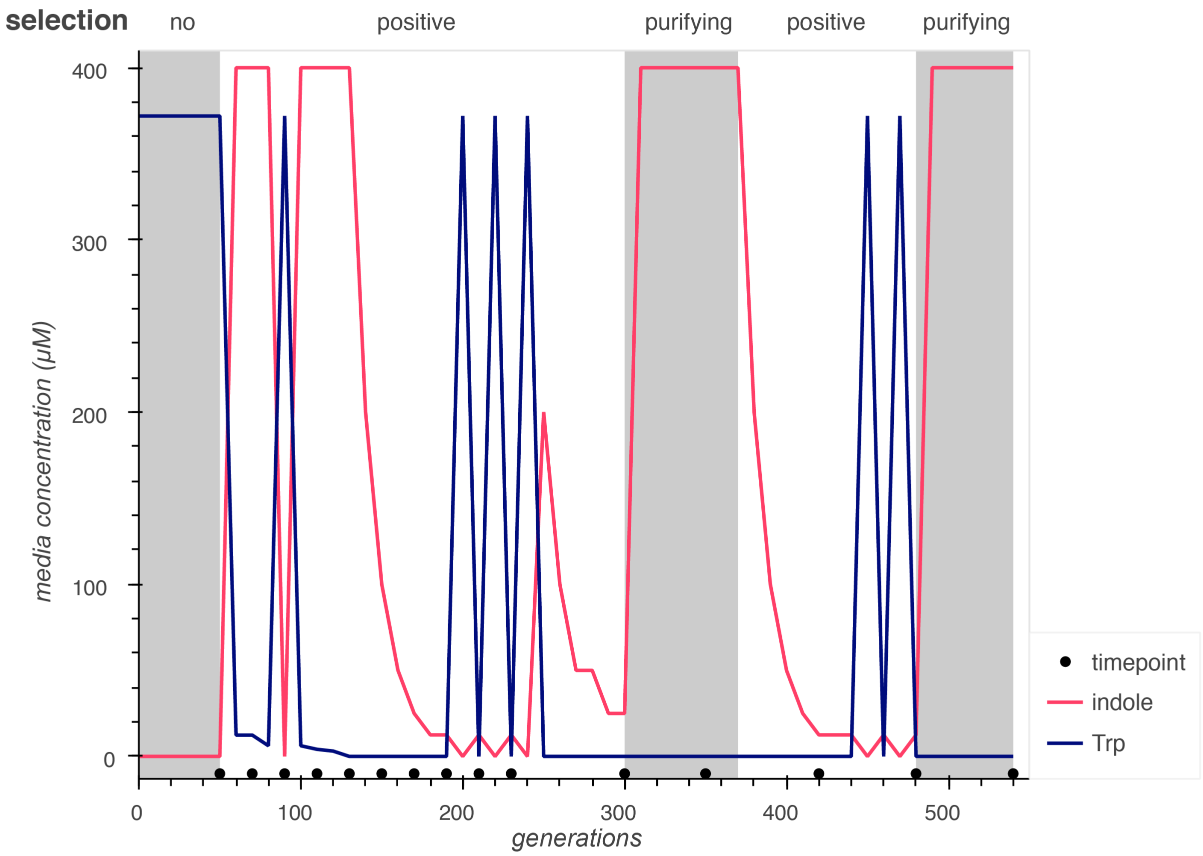


**Fig. S14. Selection condition schedule for TrpB evolution.** All 96 cultures in the experiment were passaged in synthetic complete medium with the indicated concentrations of indole and Trp, and DNA was harvested and sequenced at the indicated timepoints. Intermittent passages into Trp-supplemented media during positive selection phases were employed to increase the rate of diversification.

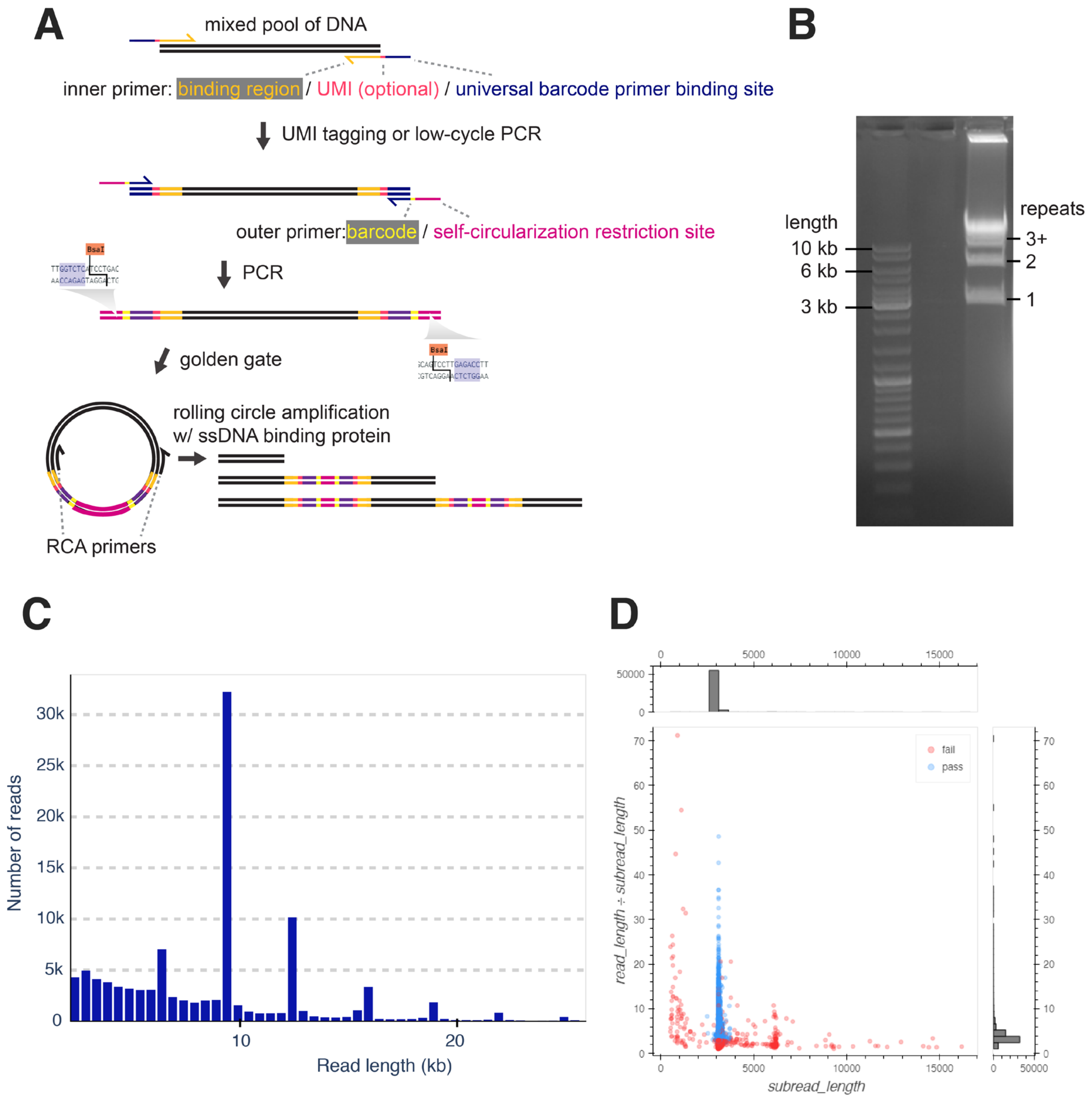


**Figure S15. A modified rolling circle amplification (RCA) method for high accuracy long read nanopore sequencing.** (**A**) Illustration of the steps involved in the RCA method. Note that RCA primers must bind internally to the inner primers to prevent amplification of primer dimers. (**B-D**) An example RCA and nanopore sequencing results using a Flongle flow cell for an amplicon of 3 kb in length, showing gel electrophoresis of the RCA product (B), the distribution of read lengths (C), and the distribution of the number of repeats (estimated as the ratio of read length to subread length) vs. subread length (D). Number of repeats and subread length were used for quality filtering prior to downstream analysis, the results of which are shown in color in D. Note that the bands corresponding to 3+ repeats were gel extracted for sequencing, resulting in the minimal amplicons with 2 or fewer repeats observed in C.

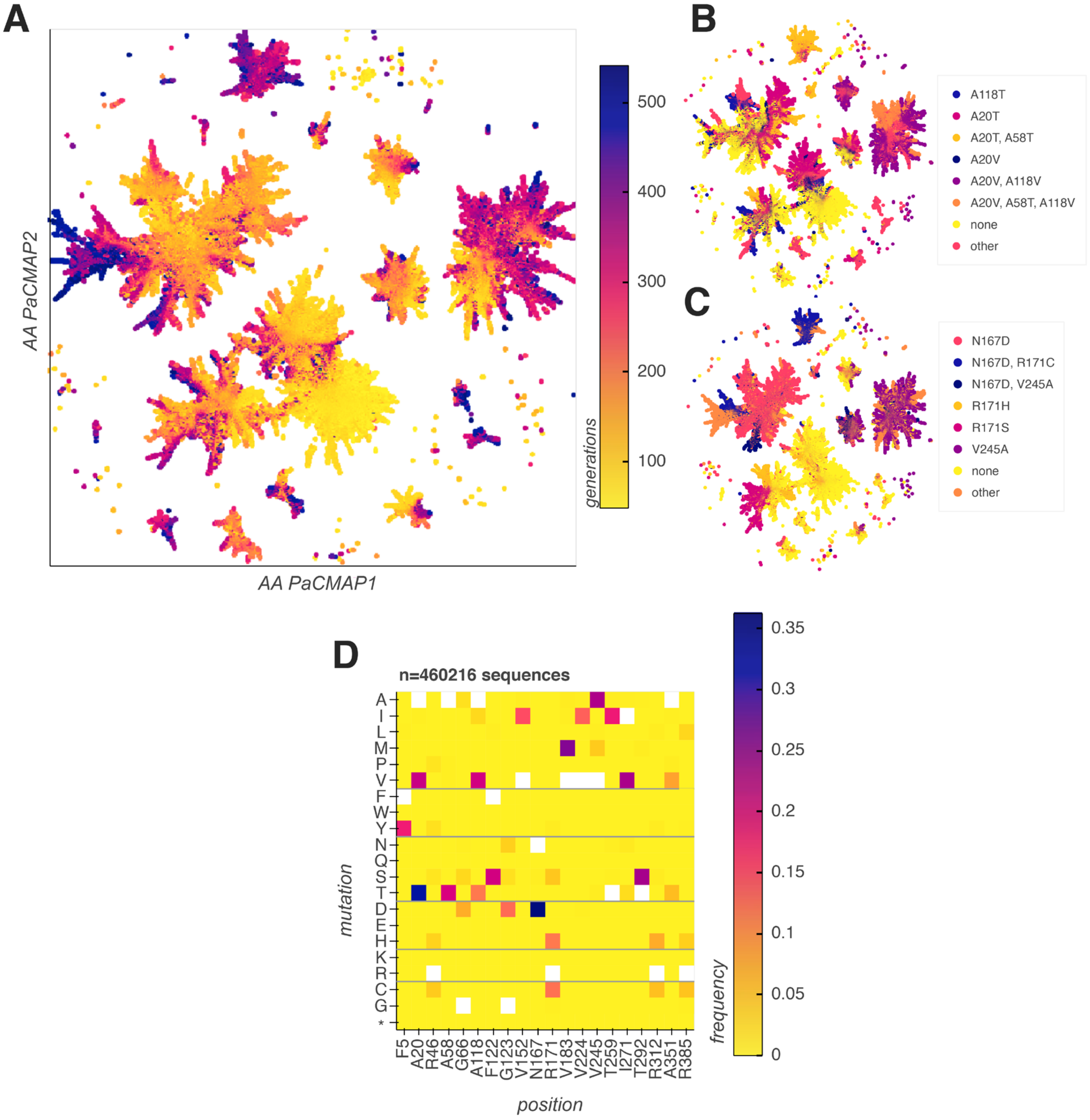


**Figure S16. Diversity of evolved TrpB variants.** (**A**) 2-dimensional representation for all unique genotypes identified using PaCMAP for dimensionality reduction, with the timepoint from which each genotype was identified represented as color. (**B** and **C**) 2D representations as in E, with color used to indicate the most frequently observed combinations of mutations among the six most frequently mutated positions in the final timepoint. (**D**) Heatmap of all mutations to the 20 most commonly mutated positions throughout the entire experiment.

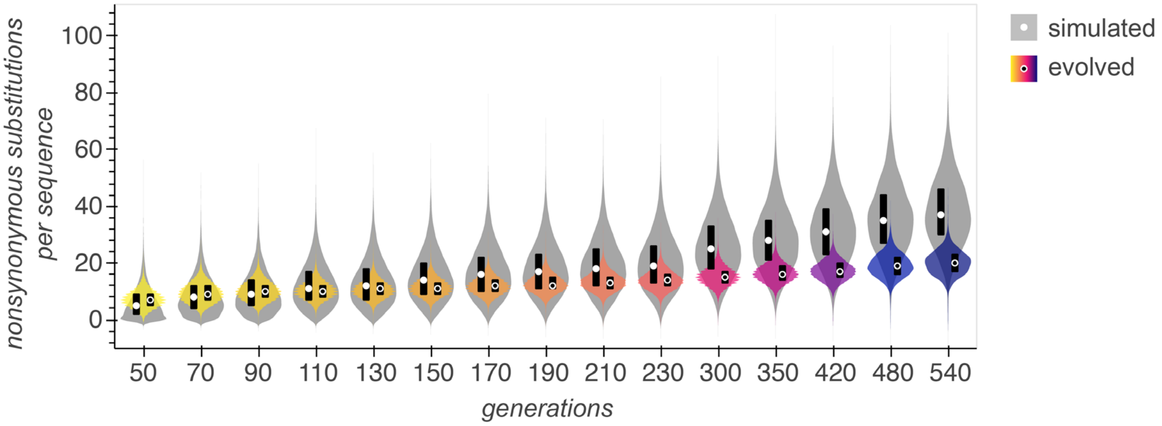


**Figure S17. Nonsynonymous mutation distributions for simulated and evolved sequences.** Violin plots of nonsynonymous mutations per sequence in each timepoint of TrpB evolution (colors) compared to that of sequences from a simulated dataset with an equivalent number of synonymous mutations generated at random using mutation rates and preferences of BadBoy2 (grey). Points and black bars denote the means and interquartile range for all sequences within each timepoint.

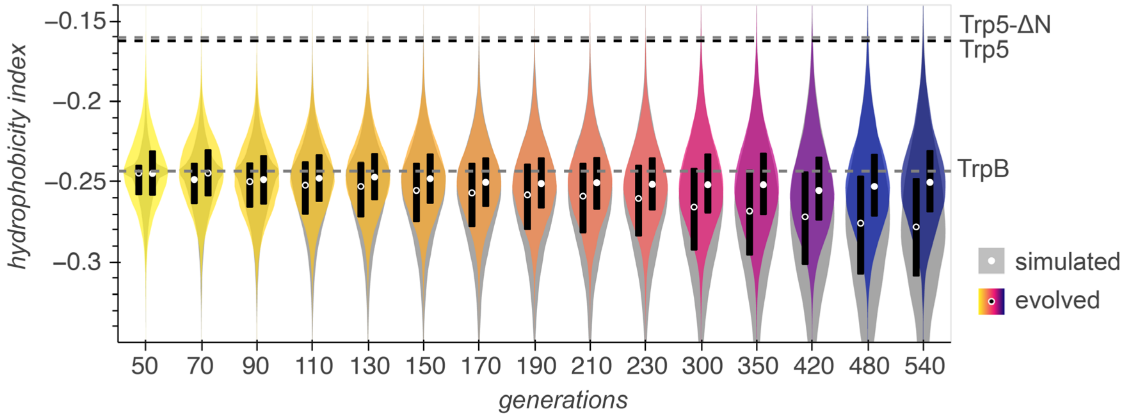


**Figure S18. Effect of long-term mutagenesis on hydrophobicity.** Violin plots of Kyte-Doolittle hydrophobicity index in each timepoint of TrpB evolution for both evolved (colors) and simulated sequences (grey). Points and black bars denote the means and interquartile range for all sequences within each timepoint. Hydrophobicity indices of the wild type TrpB, the TrpA-TrpB holoenzyme ortholog from *Saccharomyces cerevisiae* (Trp5) and an N-terminally-truncated Trp5 homologous to TrpB (Trp5-ΔN) are shown for comparison.

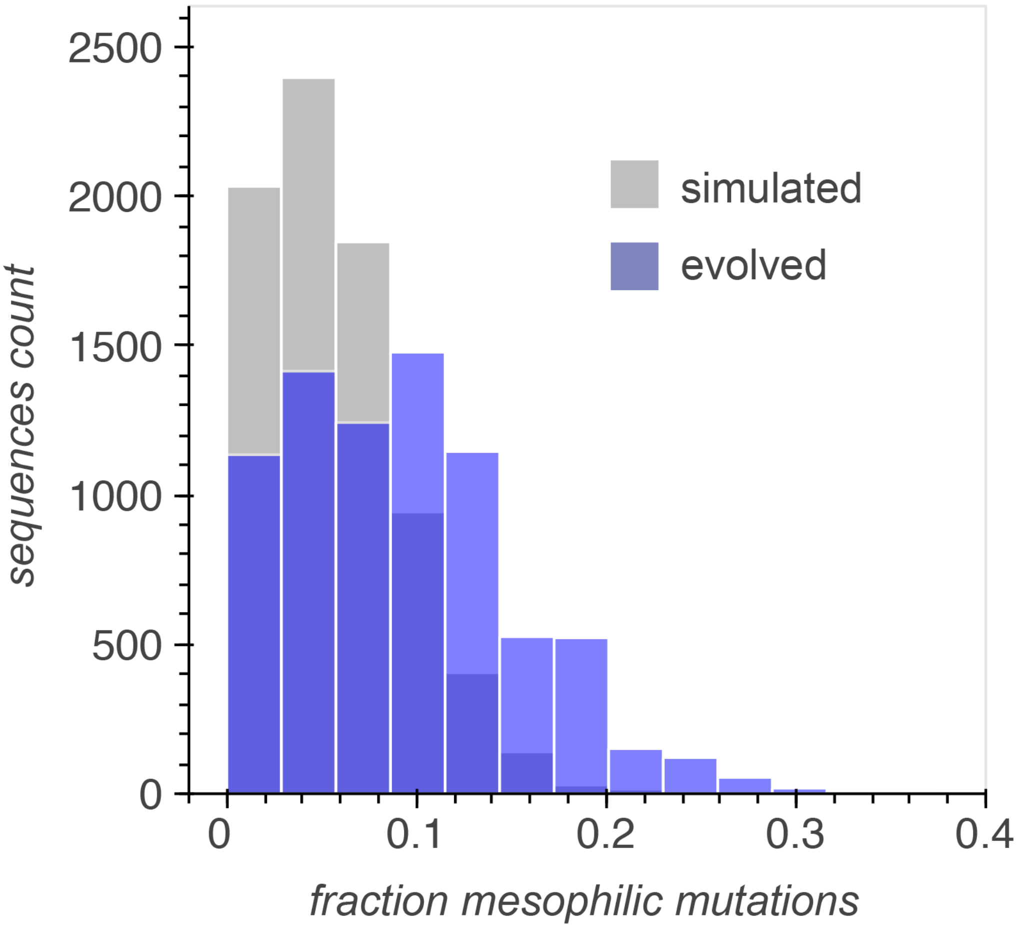


**Figure S19. Distribution of mesophile mutation fraction for both evolved and simulated TrpB sequences.** Mesophile mutation fraction was calculated for each sequence as the number of mesophile mutations divided by the total mutations. A set of 17 amino acid replacements (*e.g.* Proline to Serine) identified by Haney *et al.*[14] were designated as mesophile mutations (see Methods). Only sequences from the final timepoint (540 generations) were included.

## Supplementary Tables

| | directed evolution round | substitution rate (per base per generation) | | | | |
|---|---|---|---|---|---|---|
| **TPDNAP variant** | | **total** | **Tv** | **Ts** | **deletion** | **insertion** |
| **WT I777K-L900S** | previously published | 0.0000107 | 4.16E-07 | 1.08E-05 | 4.96E-06 | 9E-07 |
| **TKS** | epPCR 1 | 0.0000136 | 7.89E-07 | 1.31E-05 | 4.69E-06 | 2E-06 |
| **SgtKis** | epPCR 2 | 0.0000263 | 2.59E-06 | 2.37E-05 | 1.48E-06 | 1E-05 |
| **Trixy** | epPCR 2 | 0.0000393 | 2.32E-06 | 3.71E-05 | 2.01E-06 | 3E-06 |
| **BadBoy1** | recombination | 0.000116 | 1.97E-05 | 9.68E-05 | 3.12E-06 | 2E-05 |
| **BadBoy2** | recombination | 0.000144 | 2.34E-05 | 0.000121 | 3.38E-06 | 3E-05 |
| **BadBoy3** | recombination | 0.000103 | 1.69E-05 | 8.64E-05 | 3.21E-06 | 3E-05 |
| **3A** | epPCR 3 | 0.0000392 | 1.99E-06 | 3.73E-05 | 2.01E-06 | 5E-06 |
| **3B** | epPCR 3 | 0.0000274 | 1.01E-05 | 1.75E-05 | 3.4E-06 | 5E-05 |
| **3C** | epPCR 3 | 0.0000185 | 6.69E-06 | 1.19E-05 | 3.98E-06 | 3E-05 |
| **3D** | epPCR 3 | 0.000074 | 2.87E-06 | 7.11E-05 | 2.98E-06 | 7E-06 |
| **3E** | epPCR 3 | 0.0000672 | 2.15E-06 | 6.76E-05 | 3.68E-06 | 5E-06 |
| **BB-3A** | epPCR 3 + BadBoy3 | 0.0000887 | 1.26E-05 | 7.9E-05 | 9.52E-07 | 4E-05 |
| **BB-3B** | epPCR 3 + BadBoy3 | 0.0000391 | 9.79E-06 | 2.96E-05 | 4.09E-07 | 2E-05 |
| **BB-Tv** | epPCR 3 + BadBoy3 | 0.0000166 | 7.34E-06 | 9.39E-06 | 1.36E-06 | 2E-05 |
| **BB-5k** | epPCR 3 + BadBoy3 | 0.000172 | 2.62E-05 | 0.000148 | 3.57E-07 | 3E-05 |
| **BB-3E** | epPCR 3 + BadBoy3 | 0.0000986 | 1.37E-05 | 8.53E-05 | 2.28E-06 | 3E-05 |

**Table S3.1. Polymerase variants and mutation rates measured in HTS dataset 6.**

| Group | Total sequences | Mean nonsynonymous mutations per sequence | Mean nucleotide mutations per base | Mean nucleotide mutations per sequence |
|---|---|---|---|---|
| P1 | 32925 | 7.495854 | 0.008871 | 11.34 |
| P3 | 36907 | 9.645325 | 0.011602 | 14.83 |
| P5 | 39504 | 10.41041 | 0.012753 | 16.3 |
| P7 | 39341 | 10.59696 | 0.014051 | 17.96 |
| P9 | 29708 | 10.94645 | 0.01471 | 18.8 |
| P11 | 30118 | 11.67285 | 0.016253 | 20.77 |
| P13 | 28393 | 12.22178 | 0.017549 | 22.43 |
| P15 | 19388 | 13.01047 | 0.018721 | 23.93 |
| P17 | 14742 | 13.81081 | 0.019951 | 25.5 |
| P19 | 19715 | 14.31291 | 0.021 | 26.84 |
| HTS_control (P1-P19) | 8517 | 0.277797 | 0.000355 | 0.45 |
| P26 | 56928 | 15.74666 | 0.024755 | 31.64 |
| P31 | 41431 | 16.80213 | 0.026753 | 34.19 |
| P38 | 30775 | 17.84195 | 0.029298 | 37.44 |
| P44 | 18678 | 19.66988 | 0.0328 | 41.92 |
| P50 | 9154 | 20.61132 | 0.034815 | 44.49 |
| HTS_control (P26-P50) | 22771 | 0.186465 | 0.000247 | 0.32 |

**Table S3.2. Basic mutation statistics for *Tm*TrpB evolution (HTS dataset 6).**

| Plasmid ID | Plasmid name | E coli selection marker | yeast selection marker |
|---|---|---|---|
| pGR438 | pUC-LEU2int-reTmTrpB(10B2)-10xNbarcode | AmpR | leu2NTD (requires leu2CTD landing pad) |
| pGR590 | AmpR | leu2NTD (requires leu2CTD landing pad) | |
| pGR480 | AmpR | leu2NTD (requires leu2CTD landing pad) | |
| pGR481 | pUC-LEU2int-trp5.K384*-ura3.K93N | AmpR | leu2NTD (requires leu2CTD landing pad) |
| pGR475 | pCAN1int-CAN1-WT.TPDNAP.rvs(REV1p)-HygR | AmpR | HygR |
| pGR420 | pUC-pGKL1int-leu2CTD-junk1-rMet15g1(p2ORF5)-5kb | AmpR | leu2NTD (requires leu2CTD landing pad) |
| pGR562 | AmpR | LEU2NTD | |
| pGR326 | AmpR | URA3 | |
| pGR550 | pNP.UMI.seq-kanR_ccdB_PaqCI | KanR | N/A |
| pGR554 | pNP.UMI.seq-kanR_ccdB_BsaCI | KanR | N/A |
| pGR518 | AmpR | NatMX | |
| pGR595 | pCAN1int.ISceI.expr-EcoRI.site-TP-DNAP1.Badboy2(SAC6p)-NatMX | AmpR | NatMX |
| pUMI | KanR | LEU2 | |
| pGR686 | RPL18Bp-Tm1Ff-barcode-library | KanR | LEU2 |
| pGR687 | RPL18Bp-TmTriple-barcode-library | KanR | LEU2 |
| pGR685 | RPL18B-TrpB.library | KanR | LEU2 |
| pFLO1-KO | FLO1.USflank-KOflank-URA3-FLO1.DSflank | N/A | URA3 |

**Table S3.3. Plasmids used in this study.**

| primer pair | related to | amplicon | fwd primer | rvs primer |
|---|---|---|---|---|
| 1 | HTS dataset 1 | LEU2 C terminus | ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXgacatgtaaaatgtttgttcc | GACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXXXXXactatcccaagcgacac |
| 2 | HTS dataset 2/3/4/5/7 | recombinant p1 sequence | ATCGAGTCAGTGCGAGTGNNNYRNNNYRNNNYRNNNgtatgacctaattgactccg | CGTGTAGAGACTGCGTAGGNNNYRNNNYRNNNYRNNNgatgacctatacataggaagatctatag |
| 3 | HTS dataset 2/3/4/5 | Universal UMI-tagged amplicon | TGTAGCGTCACCTGCTATGGAAAGCXXXXXXATCGAGTCAGTGCGAGTG | TGACTCCACACCTGCATACAGGAGATGXXXXXXCGTGTAGAGACTGCGTAGG |
| 4 | HTS dataset 6 | LEU2-URA3-mScarlettI | ATCGAGTCAGTGCGAGTGNNNYRNNNYRNNNYRNNNcaaggattttcttaacttcttcg | CGTGTAGAGACTGCGTAGGNNNYRNNNYRNNNYRNNNgatgacctatacataggaagatctatag |
| 5 | HTS dataset 6 | Universal UMI-tagged amplicon | AAGGTCTCAAGGACACCTGCTATGGAAAGCXXXXXXXATCGAGTCAGTGCGAGTG | TTGGTCTCATCCTCACCTGCATACAGGAGATGXXXXXXXCGTGTAGAGACTGCGTAGG |
| 6 | HTS dataset 7 | Universal UMI-tagged amplicon | TTCACCTGCTATGTCCTGGTCTCAGAAAGACGXXXXXXXATCGAGTCAGTGCGAGTG | AACACCTGCATACAGGAGGTCTCAAGGACTGCXXXXXXXCGTGTAGAGACTGCGTAGG |
| 7 | HTS dataset 7 | RCA TrpB | GATGACCTATACATAGGAAGATC | gactccggcgaaaaagcatg |
| 8 | HTS dataset 8a | UMI-p1ORF4 (filler DNA) w/ BsmBI flanks | aaaacgtctcagctgaAANNYRNNYRNNYRNNYRNNYRNNYRNNYRaagtagcaccgcctaaccctctgctaatgaatttgaaggcgata | aaaacgtctctgagtactcctcagactcagattggtgtttc |
| 9 | HTS dataset 8a | TmTrpB w/ golden gate overhangs | gtcggcCACCTGCtatatATGAAAGGCTACTTCGGTC | GAGCATCATCATCATCATCATTAAatccacgaGCAGGTGAACCAA |
| 10 | HTS dataset 8b | TmTrpB genotype UMI (inner) | ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXcggcatgccgagcaaatg | TTCAGACGTGTGCTCTTCCGATCTXXXXXXXagatatcgccttcaaattcattagcagag |
| 11 | HTS dataset 8b | TmTrpB genotype UMI (outer) | AATGATACGGCGACCACCGAGATCTACACatagaggcACACTCTTTCCCTACACGACGCTCTTCCGATCT | CAAGCAGAAGACGGCATACGAGATtctccggaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC |
| 12 | epPCR 1, TP-DNAP1 engineering | TP-DNAP1 exonuclease domain (~residues 346 - 575) | aaaattttcaccattgtccaagttcag | ccaattcgcagttaaaaatatcttc |
| 13 | epPCR 1, TP-DNAP1 engineering | TP-DNAP1 motif A (~residues 564 - 775) | actgtagaaacgacgtattggtcttgtc | cacaatgagttcatgatCTttttgatg |
| 14 | epPCR 2 and 3, TP-DNAP1 engineering | TP-DNAP1 sans TP (~ residues 255 - C terminus) | cacttcggtctcaagcaatgttcaagatttttgtggc | tggactggtctcataagaaattcgCCTCGAGttagg |
| 15 | *Tm*TrpB integration cassette, TrpB evolution | *Tm*TrpB w/ lineage barcode | aaatgtagaaacatgTGATAAGCTCATAGACATGTAAA | ctaattgactccggcgaaaaagcatgcXXXXXXXXXXXXXgagctcTTAATGATGATGATG |

**Table S3.4. Primers pairs used in this study.**

| Yeast Strain | Genotype | Notes | Parent Strain | Source |
|---|---|---|---|---|
| OR-Y219 | MAT**α** *his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0* ρ+ | BY4742 | See Brachmann et al, 1998 | ATCC 201389 |
| OR-Y260 | MAT**α** *his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 trp5Δ0* ρ+ | TRP5 deletion, used for generating HTS dataset 8 | OR-Y219 | This work |
| AR-Y404 | MAT**a** *can1 his3 leu2Δ0 HIS4* | AH22 | See previous work | Previous work (Ravikumar et al 2018) |
| OR-Y484 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-HygR ρ+ + p2 + p1-LEU2CTD-junkDNA-Met15-5kb | TRP5 / FLO1 / MET15 knocked out, his3Δ1 deletion, recoded TP-DNAP1 (pGR475) integrated, p1/p2 protoplast fused, p1 landing pad (pGR420) integrated | AR-Y404 | This work |
| OR-Y487 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-HygR ρ+ + p2 + p1-LEU2-URA3-TRP5 | LEU2-URA3-TRP5 p1 (pGR480) integrated | OR-Y484 | This work |
| OR-Y488 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-HygR ρ+ + p2 + p1-LEU2-ura3*-trp5* | LEU2-ura3*-trp5* p1 (pGR481) integrated | OR-Y484 | This work |
| OR-Y516 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-HygR ρ+ + p2 + p1-LEU2-mScarlettI | mScarlettI (pGR562) p1 integration, used for generating HTS dataset 4 | OR-Y484 | This work |
| OR-Y532 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-HygR ρ+ + p2 + p1-LEU2-mScarlettI-ura3.K75* | mScarlettI-ura3-K75* (pGR590) p1 integration, used for generating HTS dataset 6 | OR-Y484 | This work |
| OR-Y538 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::TPDNAP-TrxSgk23 (pGR595) ρ+ + p2 + p1-LEU2CTD-junkDNA-Met15-5kb (pGR420) | TP-DNAP1-Badboy2 (pGR595) integration | OR-Y484 | This work |
| OR-Y539 | MAT**a** *his3Δ1 leu2Δ0 ura3Δ0 HIS4 trp5Δ0 flo1Δ0 met15Δ0* CAN1::WT-TPDNAPv3-RPL18B-HygR (pGR475) ρ+ + p2 + p1-LEU2-*Tm*TrpB (pGR438) | p1 integration, ~400 colonies harvested together | OR-Y538 | This work |

## Table S3.5. Yeast strains used in this study

| **data set label** | **experiment** | **Related to** | **library preparation method** | **p1 integration casette or relevant plasmid(s)** |
|---|---|---|---|---|
| HTS dataset 1 | Mutation accumulation with legacy TPDNAPs | Extended Data Figure 1, Extended Data Figure 2 | 1-step PCR | pGR326 |
| HTS dataset 2 | Mutation accumulation, TPDNAP error prone PCR 1 | Extended Data Figure 7 | *in vivo* downsampled UMI PCR | pGR481 (LEU2-ura3*-trp5*) |
| HTS dataset 3 | Mutation accumulation, TPDNAP error prone PCR 2 | Extended Data Figure 8 | *in vivo* downsampled UMI PCR | pGR481 (LEU2-ura3*-trp5*) |
| HTS dataset 4 | Mutation accumulation, TPDNAP recombination | Extended Data Figure 9 | *in vivo* downsampled UMI PCR | pGR562 |
| HTS dataset 5 | Mutation accumulation, TPDNAP error prone PCR 3 | Extended Data Figure 10 | *in vivo* downsampled UMI PCR | pGR481 (LEU2-ura3*-trp5*) |
| HTS dataset 6 | TPDNAP epPCR3 + recombination 4 timepoint mutation accumulation and TPDNAP sequencing | Figure 1, extended data figure 11-13 | *in vivo* downsampled UMI PCR | pGR590 |
| HTS dataset 7 | TrpB evolution with TPDNAP BadBoy2 | Figure 2, Figure 3, Figure 4 | rolling circle amplification | pGR438 |
| HTS dataset 8a | Evolved TrpB fitness assay, genotype <==> UMI | Figure 5 | *in vivo* downsampled UMI PCR | pGR438 |
| HTS dataset 8b | Evolved TrpB fitness assay, UMI <==> fitness | Figure 5 | 2-step PCR | pGR438 |

**Table S3.6. High throughput sequencing datasets generated and referenced in this study.**