Title (100 characters)

Efficiently exploring functional space in loop engineering with variations in length and composition

Authors:

Pedro A. G. Tizei[1], Emma Harris[2] , Marleen Renders[3], and Vitor B. Pinheiro[1,2*]

Affiliation:

[1] University College London, Department of Structural and Molecular Biology, Gower Street, London, WC1E 6BT, UK.

[2] Birkbeck, Department of Biological Sciences, University of London, Malet Street, WC1E 7HX, UK.

[3] Laboratory for Medicinal Chemistry, Rega Institute for Medical Research, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

*corresponding author – v.pinheiro@ucl.ac.uk

Abstract:

Insertions and deletions (indels) can affect function[1], biophysical properties[2]and substrate specificity[3] of enzymes, and they play a central role in evolution. Despite such clear relevance, this class of mutation remains an underexploited tool in protein engineering with no available platforms capable of systematically generating and analyzing libraries of varying sequence composition and length. Here, we present a novel DNA assembly platform (InDeL assembly), based on cycles of endonuclease restriction and ligation, that coupled to a *k*-mer based sequence analysis framework, enables systematic and efficient navigation of the sequence space across different compositions and lengths. We demonstrate the approach by engineering the well-characterized TEM-1 β-lactamase Ω-loop, involved in substrate specificity, identifying novel extended spectrum β-lactamases in areas of the sequence space not previously explored. InDel assembly provides a route to optimize protein loops or linkers where sequence length and composition are both essential functional parameters.

Directed evolution is a powerful tool for optimizing, altering or isolating novel function in proteins and nucleic acids[4, 5]. Through cycles of sequence diversification and purifying selection, it allows a given sequence space to be explored in search of a desired function - with library quality and diversity as factors on how efficiently that search can be carried out and on how much of the available sequence space can be explored.

Diversity can be introduced into a gene of interest by a number of well-established strategies - that vary in cost, in how that diversity is distributed and in the level of customization that can be implemented[6, 7]. PCR-based methods using degenerate primers provide a cost-effective route towards creating focused (i.e. that target a small number of clustered sites) high-quality libraries[8] but cannot rival commercial high-throughput DNA synthesis platforms[9], or DNA assembly methods that rely on the incorporation of individual triplets[10, 11], for customization. While successful, these approaches focus on generating libraries of constant length, capable of efficiently sampling a given sequence landscape (of fixed-length) but unable to explore the entire available sequence space, i.e. landscapes of different lengths.

This is particularly relevant to the engineering of systems in which loops contribute significantly to protein function, such as the H3 loop in antibodies[12] or the loops in TIM barrel enzymes[3]. In those circumstances, methods that target loop composition as well as length are essential for efficient functional optimization. Traditional methods can address the problem by brute force, sampling sequence space through the use of multiple libraries, each with a given length[13].

Methods for library assembly that exploit changes in length have been developed and they successfully address the challenge of making deletions (or insertions) that do not change the reading frame of the targeted gene - be it through modifying oligo synthesis[14], by using insertion and excision cycles of engineered transposons[15] or by combining chemical and enzymatic approaches[16]. Nevertheless, they have had limited impact in the engineering of protein loops: Aside from the generation of diversity, analysis of selection outputs that vary in length as well as composition is challenging. Selection can be carried out until population diversity is sufficiently low that analysis is redundant, or by analyzing single-length landscapes[17]. Either approach sacrifices significant evolutionary information that is relevant to protein engineering.

Here, we present a combination of (i) a cost-effective DNA assembly of high-quality, highly customizable focused libraries capable of sampling both length and compositional variation, and (ii) a robust analytical framework that enable the analysis of selection outputs that include both length and compositional variation. Together, they make an ideal platform for efficiently navigating sequence space and for engineering protein loops and linkers.

Analogous to Sloning[10] and ProxiMAX[11], InDel assembly relies on cycles of restriction and ligation, using type IIs endonucleases and custom dsDNA building blocks, to assemble a library on a solid support (Fig. 1a). While both commercial platforms make use of steps to ensure only full-length libraries are carried forward, our InDel platform exploits that restriction and ligation reactions are not carried out to completion in the experimental setup. The result is a combinatorial library where individual building blocks may or not be incorporated with each cycle (Fig. 1c) - reminiscent of COBARDE[14].

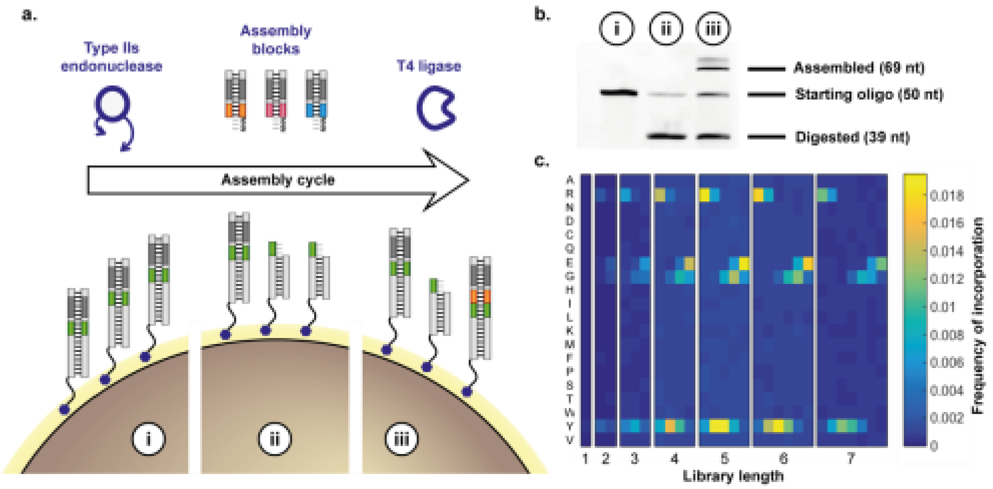


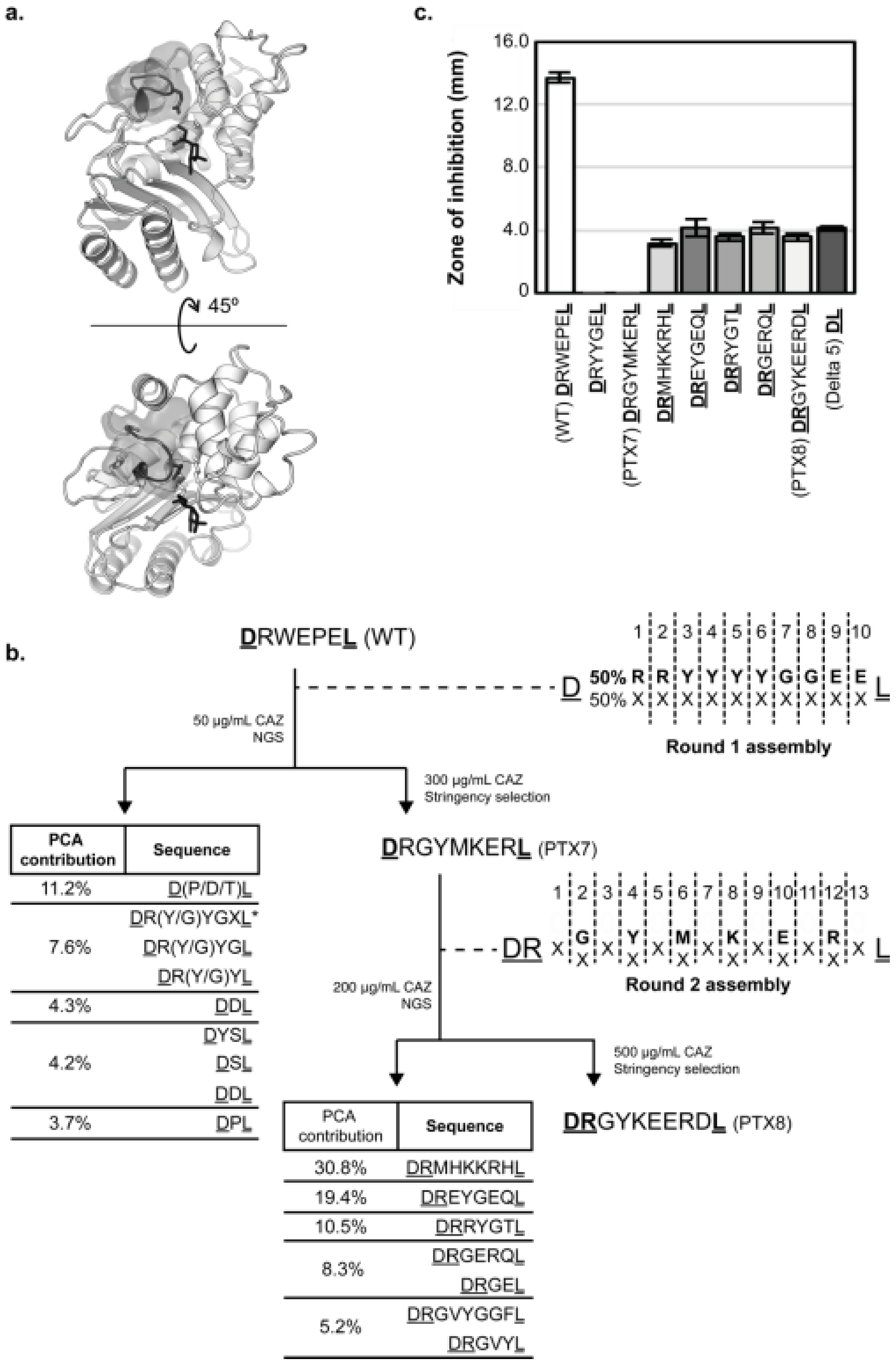
Figure 1 Producing length and compositional variation with InDel assembly. (**a**) Each assembly cycle starts with dsDNA templates bound to paramagnetic beads (i) that are restricted with type IIs endonucleases (ii). DsDNA assembly blocks are annealed and ligated (iii) to complete the cycle. (**b**) Denaturing gel electrophoresis of bead-bound oligos after steps (i-iii), detected by a fluorescent label in the template oligo. (**c**) Expected library composition (1st round library depicted in Fig. 2b) after 10 rounds of assembly (at 50% cycle efficiency) enriching for the desired RYYGE motif in the 5 amino acid library. Libraries longer than 7 are omitted for clarity.

Using a fluorescently labelled initiation block, we optimized ligation (Supplementary Fig. 1) and restriction conditions, explored building block topology (i.e. hairpins or dsDNA from annealed strands), explored creating library degeneracy through concomitant addition of multiple blocks, and other reaction parameters. Gel-based assays suggested 50% assembly efficiency per cycle could be obtained (Fig. 1b), however sequence analysis of synthesized libraries determined significantly lower efficiencies (Supplementary Fig. 2), probably the result of extended sample handling. Codon bias was observed but varied between libraries assembled.

Having established the assembly platform, we chose the β-lactamase TEM-1 to demonstrate its potential. TEM-1 is a well-characterized enzyme[18, 19], with a long track record in directed evolution through its ease of selection, wide range of available substrate analogues and its clinical relevance[20, 21]. In particular, TEM-1 contains a short flexible loop which is part of the active site (the Ω-loop, ${}_{164}$RWEPE${}_{168}$ - Fig. 2a) and that has been implicated in substrate specificity. Palzkill and colleagues[22, 23] have systematically explored the sequence landscape of the Ω-loop to isolate enzymes with a different substrate spectrum - altering TEM-1 resistance from the penam ampicillin (AMP) to the cephem ceftazidime (CAZ). In addition, even though their libraries were designed to explore only the sequence landscape of similar length to wild-type TEM-1, we were encouraged they also reported shorter variants with significant CAZ resistance[22].

Our initial goal was to explore the sequence neighborhood of the previously reported ${}_{164}$RYYGE${}_{168}$ variant and we assembled a library heavily biased towards generating it (Fig. 2b). Based on our early estimates of 50% efficiency of assembly per cycle, we carried out 10 cycles of assembly using a degenerate mixture of building blocks - 50% coding for the target residue and the remaining 50% equally divided between the other 19 coding triplets.

Assembled libraries were cloned into a TEM-1 backbone harboring the M182T stabilizing mutation[20]. Selection was carried out by plating transformants harboring the TEM-1 library directly on media supplemented with 50 μg/μL CAZ - below the MIC (minimal inhibitory concentration) for the RYYGE variant harboring the stabilizing mutation (Supplementary Fig. 5). As expected, the RYYGE variant was present in the initial library (5 reads in 2.3 million - 0.0002% of the population) and was significantly enriched on selection (314 reads in 230 thousand - 0.14%) - an enrichment score (Z-score based on comparison of two Poisson distributions) of 8460 (in the 99th percentile of the Z-score distribution). Because of the high number of transformants obtained in selection, we decided to increase selection stringency to isolate only the most active variants. Plating the library at high CAZ concentrations (300 μg/μL), only a handful of transformants could be isolated. Sequencing confirmed that all colonies harbored the same variant: RGYMKER (labelled PTX7).



PCA contribution	Sequence
11.2%	D(P/D/T)L
7.6%	DR(Y/G)YGXL* DR(Y/G)YGL DR(Y/G)YL
4.3%	DDL
4.2%	DYSL DSL DDL
3.7%	DPL

PCA contribution	Sequence
30.8%	DRMHKKRHL
19.4%	DREYGEQL
10.5%	DRRYGTL
8.3%	DRGERQL DRGEL
5.2%	DRGVYGGFL DRGVYL

Figure 2 Directed evolution of TEM-1 Ω-loop variants with altered substrate specificity. (**a**) TEM-1 beta-lactamase in complex to substrate analogue inhibitor (PDB: 1TEM). Inhibitor is shown in the TEM-1 active site in black, and side chains of key catalytic residues (S70, K73 and E166) are shown. The Ω-loop is shown in grey and superimposed to its translucent space-filling

representation. (**b**) Summary of the directed evolution of TEM-1. A 1st round library, biased towards the generation of RYYGE variant was assembled and selected. Next generation sequencing (NGS) confirms RYYGE was significantly enriched (**D**RYYGE**L** was the 16th most frequent sequence post-selection and picked up in the 2nd PCA dimension - see Supplementary Table 1). PTX7, isolated from a high-stringency selection, was used as seed for the 2nd round library. PTX8 and sequences from the top four PCA dimensions were further characterized. (**c**) Ceftazidime resistance of selected variants, measured by inhibition of growth around antibiotic-soaked paper discs - higher values indicate lower resistance (n = 3, error bars represent s.e.m). Underlined residues represent the invariant edges of the assembly, constant in each library and required for library amplification.

Undetected in the input library, PTX7 represented approximately 0.004% of the selected library (9 reads in 230 thousand). Crucially, PTX7 differs from WT (RWEPE) and engineered TEM-1 (RYYGE) in both composition and length. This highlights the power of InDel to navigate the sequence space since loops longer than wild-type would not normally be investigated and a brute force approach would be seeking a sequence among 1.2 x 10^9 (assuming all 7-mer motifs were represented once in this imaginary library).

It was clear that the functional density of the sequence space around TEM-1 Ω-loop was higher than previously reported. We therefore decided to pursue a second round of selection to investigate the sequence space in the neighborhood of PTX7, including deletions, substitutions and insertions. Exploring that sequence space could easily be achieved with InDel by assembling a library alternating between fully degenerate (i.e. equal distribution of all available triplets) with biased (i.e. 50% of desired PTX7 triplet and 50% of the remaining 19 triplets) cycles (Fig. 2b, Supplementary Fig. 3 and Supplementary Fig. 4).

Selection was carried out by plating the newly assembled libraries on high concentrations of CAZ (200 µg/µL), from which 79 colonies were isolated. Recovered transformants were further selected at higher selection stringency (500 µg/µL) with a single variant being identified: RGYKEERD (labeled PTX8). This variant was detected in the input library for the second round selection (1 read in 6.5 million) and enriched significantly upon selection (550 reads in 64 thousand) - a Z-score of 5914 (in the 95th percentile).

Recloning PTX8 onto a naive background, demonstrated that PTX8 was in fact a selection parasite. Although it is more resistant to CAZ than wild-type TEM-1 (Fig. 2c), a mutation crept into the plasmid origin of replication, increasing plasmid copy number per cell and likely expression levels of the mutant β-lactamase.

Use of deep sequencing to map functional landscapes[24] and to accelerate directed evolution[25] is well-established, but do not perform well for short libraries that vary in both length and composition - length variants are generally discarded from the analysis[26]. We therefore considered whether alignment-free sequence analysis methods based on subdivision of sequence strings into short reading windows (*k*-mers) could be adapted to extract information from sequencing data spanning multiple fixed-length sequence landscapes.

K-mer based methods are integral components of large sequence comparison methods[27] as well as next generation sequence assembly[28], allowing comparison of sequences of different lengths as well as reconstruction of sequence motifs - the two steps required to identify functional variants from the available InDel assembly libraries.

We opted for using masked 3-mers (i.e. $X_1X_2X_3$ being considered as $X_1X_2\_$ and $X_1\_X_3$[29]) to analyze sequences, reducing computational burden (reducing the possible 8000 3-mers to 800 possible masked 3-mers) without significant loss of information relevant for motif reconstruction. Based on analysis of 'toy' data sets (not shown), we chose to explicitly represent residues flanking the synthesized libraries (as Z_1 and Z_2 - see Fig. 3a) in analysis, adding a further 82 possible *k*-mers but significantly improving the robustness of downstream motif assembly.

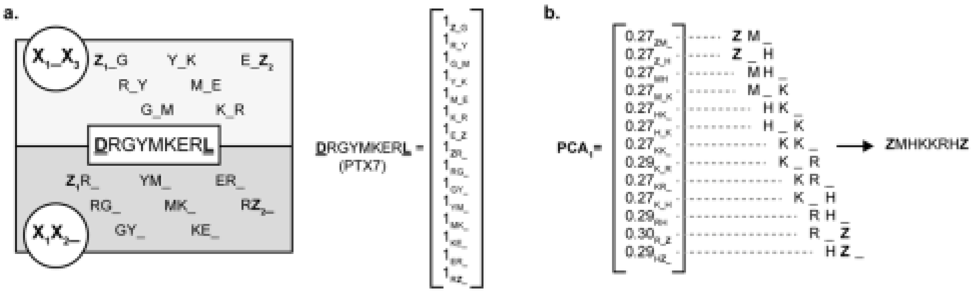

a.
X₁_X₃
Z₁_G
Y_K
E_Z₂
R_Y
M_E
G_M
K_R
DRGYMKERL
Z₁R_
YM_
ER_
RG_
MK_
RZ₂_
GY_
KE_
X₁X₂_
DRGYMKERL =
(PTX7)
1Z_G 1R_Y 1G_M 1Y_K 1M_E 1K_R 1E_Z 1ZR_ 1RG_ 1GY_ 1YM_ 1MK_ 1KE_ 1ER_ 1RZ_
b.
PCA₁=
0.27ZM_ 0.27Z_H 0.27MH_ 0.27M_K 0.27HK_ 0.27H_K 0.27KK_ 0.29K_R 0.27KR_ 0.27K_H 0.29RH_ 0.30R_Z 0.29HZ_
Z M _
Z _ H
M H _
M _ K
H K _
H _ K
K K _
K _ R
K R _
K _ H
R H _
R _ Z
H Z _
ZMHKKRHZ

Figure 3 *K*-mer sequence decomposition and reconstruction. (**a**) Sequences can be decomposed into all possible masked 3-mers (i.e. $X_1X_2X_3$ separated into $X_1X_2\_$ and $X_1\_X_3$) as shown for PTX7. Each masked 3-mer is counted generating an 882-dimension vector (non-zero elements shown). Vectors are normalized (multiplied by $\sqrt{15}$ in the case of PTX7) and multiplied by their enrichment score. Principal component analysis (PCA) identifies enriched *k*-mers, which allow a sequence to be reconstructed. An arbitrary cut-off (0.1)

can be used to minimize noise and facilitate assembly. (**b**) The sequence reconstruction of the first PCA component calculated from the second library selection.

Individual sequences, described by their masked 3-mer count, were treated as individual column vectors (Fig. 3a), normalized, and scaled by their Z-score as a measure for enrichment. Variation in such 882-dimensional representation of library (or selection output) can be deconvoluted by principal component analysis (PCA), identifying *k*-mers that contribute the most to enrichment (Fig. 3b) and enabling putative functional sequences to be reconstructed along individual PCA dimensions. As expected, highly enriched and abundant sequences contribute significantly to library variation and are identified in the lower PCA dimensions. Nevertheless, crucially, this approach can identify motifs across different lengths as well as positions that are not expected to contribute to function (Supplementary Table 1 and Supplementary Table 2).

We applied this approach to both selection libraries (Fig. 2b) and selected the top four candidates from the second round library (**D**RMHKKRH**L**, **D**REYGEQ**L**, **D**RRYGT**L**, **D**RGERQ**L**) for further characterization (Fig. 2c and Supplementary Fig. 5). All characterized variants, which include loops of five to eight amino acids long (as well as the loopless variant Δ5), are significantly more resistant than wild-type TEM-1, highlighting that the sequence space in the vicinity of the Ω-loop is populated with functional variants in multiple landscapes.

Our results provide further evidence that loops are highly evolvable[30] and also highlight how directed evolution of protein loops must take into account sequence spaces that straddle more than a single landscape. We show that the combination of InDel assembly and *k*-mer-based analysis provide a powerful framework for navigating sequence space that is not otherwise accessible. In addition, although we present here an example of InDel assembly with triplets, which is ideal for generating amino-acid-steps in libraries of protein coding genes, the platform is compatible with building blocks of mixed length, enabling a vast host of combinatorial possibilities that could be applied to the directed evolution of nucleic acid aptamers, gene expression regulatory elements and fragment-based protein engineering.

Acknowledgements

PAGT acknowledges CAPES foundation support (fellowship BEX 8985-13-8). VBP acknowledge support by the European Research Council [ERC-2013-StG project 336936 (HNAepisome)] and by the BBSRC (grant BB/K018132/1).

Author contributions

PAGT and VBP conceived the assembly scheme. MR carried out screening of DNA ligases. PAGT developed the assembly platform and carried out TEM-1 selections. PAGT and VBP conceived the analysis strategy. VBP wrote the analysis algorithm. PAGT carried out the sequence analysis. PAGT and EH carried out the characterization of isolated TEM-1 variants. PAGT and VBP wrote the manuscript.

References

1. Lin, L. et al. Improved catalytic efficiency of endo-beta-1,4-glucanase from Bacillus subtilis BME-15 by directed evolution. *Appl Microbiol Biotechnol* **82**, 671-679 (2009).
2. Arpino, J.A., Reddington, S.C., Halliwell, L.M., Rizkallah, P.J. & Jones, D.D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22**, 889-898 (2014).
3. Afriat-Jurnou, L., Jackson, C.J. & Tawfik, D.S. Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry* **51**, 6047-6055 (2012).
4. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat Rev Drug Discov* **16**, 181-202 (2017).
5. Packer, M.S. & Liu, D.R. Methods for the directed evolution of proteins. *Nat Rev Genet* **16**, 379-394 (2015).
6. Shivange, A.V., Marienhagen, J., Mundhada, H., Schenk, A. & Schwaneberg, U. Advances in generating functional diversity for directed protein evolution. *Curr Opin Chem Biol* **13**, 19-25 (2009).
7. Tee, K.L. & Wong, T.S. Polishing the craft of genetic diversity creation in directed evolution. *Biotechnol Adv* **31**, 1707-1721 (2013).
8. Tang, L. et al. Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *Biotechniques* **52**, 149-158 (2012).
9. Tiller, T. et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* **5**, 445-470 (2013).
10. Van den Brulle, J. et al. A novel solid phase technology for high-throughput gene synthesis. *Biotechniques* **45**, 340-343 (2008).
11. Ashraf, M. et al. ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochem Soc Trans* **41**, 1189-1194 (2013).
12. Tsuchiya, Y. & Mizuguchi, K. The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci* **25**, 815-825 (2016).
13. Knappik, A. et al. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* **296**, 57-86 (2000).
14. Osuna, J., Yanez, J., Soberon, X. & Gaytan, P. Protein evolution by codon-based random deletions. *Nucleic Acids Res* **32**, e136 (2004).

15. Jones, D.D. Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. *Nucleic Acids Res* **33**, e80 (2005).

16. Murakami, H., Hohsaka, T. & Sisido, M. Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat Biotechnol* **20**, 76-81 (2002).

17. Ravn, U. et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**, 99-110 (2013).

18. Jacoby, G.A. & Medeiros, A.A. More extended-spectrum beta-lactamases. *Antimicrob Agents Chemother* **35**, 1697-1704 (1991).

19. Salverda, M.L., De Visser, J.A. & Barlow, M. Natural evolution of TEM-1 beta-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol Rev* **34**, 1015-1036 (2010).

20. Kather, I., Jakob, R.P., Dobbek, H. & Schmid, F.X. Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *J Mol Biol* **383**, 238-251 (2008).

21. Dellus-Gur, E., Toth-Petroczy, A., Elias, M. & Tawfik, D.S. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J Mol Biol* **425**, 2609-2621 (2013).

22. Palzkill, T., Le, Q.Q., Venkatachalam, K.V., LaRocco, M. & Ocera, H. Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of beta-lactamase. *Mol Microbiol* **12**, 217-229 (1994).

23. Petrosino, J.F. & Palzkill, T. Systematic mutagenesis of the active site omega loop of TEM-1 beta-lactamase. *J Bacteriol* **178**, 1821-1828 (1996).

24. Stiffler, M.A., Hekstra, D.R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 beta-lactamase. *Cell* **160**, 882-892 (2015).

25. Woldring, D.R., Holec, P.V., Zhou, H. & Hackel, B.J. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* **10**, e0138956 (2015).

26. Pitt, J.N. & Ferre-D'Amare, A.R. Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376-379 (2010).

27. Gardner, S.N. & Hall, B.G. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One* **8**, e81760 (2013).

28. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).

29. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19**, 513-523 (2003).

30. Toth-Petroczy, A. & Tawfik, D.S. The robustness and innovability of protein folds. *Curr Opin Struct Biol* **26**, 131-138 (2014).

Online Methods

Assembly

All oligos used in InDel assembly were commercially synthesized (Integrated DNA Technologies). Assembly block oligos providing the 5'-end for ligation with the dsDNA template were phosphorylated in 100 µl reactions (1 nmol oligo per reaction) containing 1x NEB T4 DNA ligase reaction buffer and 1 µl NEB T4 polynucleotide kinase. Reactions were carried out for 3 h at 37˚C, followed by inactivation at 80˚C for 20 minutes. Oligos were phenol-chloroform extracted, ethanol precipitated, resuspended in 90 µl annealing buffer (10 mM Tris-HCl pH 8.0, 20 mM NaCl, 1 mM $MgCl_2$, 0.01% Tween20) and annealed to 1 nmol of the complementary assembly block strand. Building blocks coding for different amino acids were mixed post annealing to create the desired incorporation proportions.

In parallel, 60 µl of MyOne C1 streptavidin-coated paramagnetic beads (Thermo Fisher Scientific) were washed twice in BWBS (5 mM Tris-HCl pH7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween20) and incubated at room temperature (in BWBS) for 30 min in a rotating incubator, to reduce background binding. After washing, 10 pmol of biotinylated dsDNA template oligos were added to the beads and incubated overnight at room temperature in a rotating incubator. Beads were washed in BWBS and transferred to a 0.5 ml microcentrifuge tube for assembly.

Bead-bound templates were digested with *Sap*I (NEB) in 100 µl reactions (10 µl 10x CutSmart buffer, 2 µl *Sap*I, 1 µl 1% Tween20) for 2 h at 37˚C with vortexing every 15-20 minutes to keep beads in suspension. Beads were isolated and washed once in BWBS. The supernatant containing *Sap*I, was retained and stored at 4˚C for subsequent assembly cycles.

The desired mixture of building blocks was added to the washed beads, incubated at 37˚C for 30 s, followed by an additional 30 s incubation at 4˚C. Supernatant containing the building blocks was removed and beads transferred to a ligation reaction. Ligations were carried out in 100 µl reactions (10 µl T4 DNA Ligase buffer, 12 µl 1,2-propanediol, 10 µl 30% PEG-8000, 1 µl T4 DNA Ligase, 1 µl 1% Tween20, 65 µl ddH_2O) at 25˚C for 1 hour, with vortexing every 15-20 minutes.

Beads were isolated, washed in BWBS and could then be taken to start a new assembly cycle. The supernatant containing the ligase reaction mixture was retained and stored at 4˚C for subsequent cycles.

The final assembly cycle used a modified dsDNA assembly block (a 3' cap block) containing a priming site used for post-assembly library amplification. After ligation of the capping oligo, beads were resuspended in 50 µl BWBS for PCR amplification.

Denaturing polyacrylamide gel electrophoresis

Assembly reactions carried out with FAM-labelled templates could be visualized after separation by denaturing PAGE. Gels were 15% acrylamide (19:1 acrylamide:bis-acrylamide) with 8 M urea in 1x TBE. An equal volume of loading buffer (98% formamide, 10 mM EDTA, 0.02% Orange G) was added to FAM-labelled templates, and sampled were incubated at 95˚C for 5 min before being loaded onto the gel. Gels were run at a constant current of 30 mA for 1.5-2 h. FAM-labeled oligos were detected by imaging on a Typhoon FLA 9500 scanner (GE Life Sciences).

Library amplification and cloning

Assembled libraries were PCR amplified from beads in 50 µl reactions using 10 U MyTaq HS polymerase (Bioline), 0.2 µM each of oligos TEM1-InDel-AmpF(1/2, for corresponding rounds of selection) and TEM1-InDel-AmpR, 1 µl resuspended bead slurry from the assembled library, 1X MyTaq reaction buffer, and 1X CES enhancer solution[31]. Library amplifications were carried out with a 1 min denaturation at 95˚C, followed by 20 cycles of 15 s at 95˚C, 15s at 55˚C, 30 s at 72˚C, ending with a 2 min final extension at 72˚C. PCR cycles were limited to 20 in library amplifications to minimize amplification biases and reduce likelihood of secondary mutations. Multiple reactions were carried out in parallel to ensure sufficient material for cloning could be generated and the oligos harbored BsaI overhangs for seamless DNA assembly.

Vector backbones were generated by iPCR in 50 µl reactions using 1 U Q5 Hot Start DNA Polymerase (NEB), 0.2 µM each of oligos Vec-TEM1-InDel-F and Vec-TEM1-InDel-R(1/2, for corresponding rounds of selection), 1 ng pTEM1-Cam vector template, 200 µM dNTPs, 1X Q5 reaction buffer, and 1X CES enhancer solution[31]. Vector amplifications were carried out with a 30 s denaturation at 98˚C, followed by 30 cycles of 10 s at 98˚C, 20 s at 68˚C, and 1.5 mins at 72˚C, ending with a 2 min final extension at 72˚C. Multiple reactions were carried out in parallel to ensure sufficient material for cloning could be generated and the oligos harbored BsaI overhangs for seamless DNA assembly.

PCR products were purified using NucleoSpin Gel and PCR Cleanup columns (Macherey-Nagel). Purified vector DNA (5 µg) and library (1 µg) DNA were digested with *Bsa*I (NEB) and *Dpn*I (NEB) for 3 h at 37˚C in multiple parallel 100 µl reactions and again purified. Vector and library were ligated (1:3 molar ratio, 1 µg total DNA) with NEB T4 DNA ligase for 2 min at 37˚C, followed by 6 h at 25˚C and overnight at 16˚C. DNA was isolated by phenol-chloroform and ethanol-precipitated. Ligated DNA was resuspended in 5 µl ddH$_2$O and transformed by electroporation into NEB 10-beta cells.

Selection

Transformed libraries were plated on LB medium supplemented with suitable ceftazidime concentrations for selection, and incubated at 37˚C overnight. Colonies were harvested with a cell scraper, transferred to 10 ml LB medium containing ceftazidime, and incubated at 37˚C for 2-3 h. The liquid culture was split in three aliquots. One was supplemented with glycerol [to a final 20% (v/v) concentration], and flash frozen for -80˚C storage. A second was plated on LB medium containing higher ceftazidime concentrations to isolate the most active TEM-1 variants. The remainder was used for plasmid extraction.

Antibiotic susceptibility assays

The substrate spectrum of TEM-1 variants was tested by measuring the minimum antibiotic concentration that could inhibit bacterial growth in liquid culture (MIC) and by measuring the growth inhibition of bacteria on solid media. *E. coli* harboring TEM-1 variants were tested for their susceptibility against ampicillin (AMP), carbenicillin (CBN), ceftazidime (CAZ), cefotaxime (CTX) and imipenem (IMP).

For MIC determination, approximately 100 CFU (based on the dilution of a liquid culture in mid-log growth) were added to 200 µl LB medium supplemented with different antibiotic concentrations and allowed to grow overnight at 37˚C with shaking. MIC assays were carried out in 96-well flat bottom plates (Greiner). Cells were resuspended by mixing with a multichannel pipette and bacterial growth estimated from OD$_{600}$ measurements. No antibiotic controls were used to estimate the maximum growth of each strain in the experimental conditions and normalize OD$_{600}$ between independent experiments. Growth inhibition assays in liquid cultures were carried out in triplicate with the lowest concentration of the antibiotic to fully inhibit bacterial growth defining MIC for that strain.

Growth inhibition of the *E. coli* strains in solid medium was carried out using by placing filter paper discs (Oxoid) containing a known amount of each antibiotic onto a lawn of approximately 10^7 CFU. Antibiotic susceptibility was measured as the radius of growth inhibition around the antibiotic disc. At least three independent experiments were carried out for each strain.

DNA library preparation for next generation sequencing (NGS)

Libraries for Illumina MiSeq sequencing were prepared by PCR with oligos containing required adaptors and unique indices to allow all pre- and post-selection libraries to be sequenced in a single experiment.

Pre-selection libraries were amplified directly from the streptavidin beads isolated from assembly. Post-selection libraries were amplified from purified plasmid DNA extracted from recovered transformants. Libraries were amplified in 50 µl reactions using NEB Q5 Hot-start DNA polymerase to minimize amplification errors and PCR cycles capped at 20 to minimize amplification biases. Reactions contained 1 U polymerase, 0.2 µM each of oligos xxx-MiSeqF (separate oligo for each library, with varying index sequences for demultiplexing, names and sequences are in Supplementary Table 3) and TEM1-MiSeq-R, 1 ng plasmid template or 1µl resuspended bead slurry from the assembled library, 200 µM dNTPs, 1X Q5 reaction buffer, and 1X CES enhancer solution[31]. Product size and purity were checked on agarose gels and correct amplicons excised and purified using Monarch Gel Extraction (NEB).

Libraries were quantified by fluorimetry using a Qubit 3.0 (Thermo Fisher Scientific) with a dsDNA HS assay kit and pooled proportionally to obtain the desired number of reads for each sample. Sequencing was carried out on an Illumina MiSeq instrument by UCL Genomics using a 150 cycle v3 kit.

NGS data handling

Sequencing data was treated as described in Supplementary Note 1. Briefly, sequences were filtered for quality, trimmed to keep only the diversified regions, translated into protein sequences, counted, and formatted to serve as input for the *k*-mer analysis.

NGS analysis

Sequencing was modelled as Poisson distributions, to allow different populations to be compared and enrichment of individual sequences determined. All analyses were carried out in MATLAB (MathWorks). A Z-score, defined in [1] was used as a measure of comparison between pre- and post-selection distributions.

$$Z = \frac{(cX - Y) - (c\theta_X - \theta_Y)}{\sqrt{c^2\theta_X + \theta_Y}} \qquad [1]$$

Where c is the ratio in size between post- and pre-selection libraries (to correct for sampling), X is the number of counts for a test sequence in the post-selection library, Y is the number of counts for the same sequence in the pre-selection library. θ_X and θ_Y are the estimated Poisson parameters (counts are a fraction of the total reads) for post- and pre-selection libraries respectively. Z-scores give a measure of enrichment, with extremely positive values identifying the sequences most enriched.

Each sequence was decomposed into all possible masked 3-mers and the library termini encoded as “***Z***” characters (to avoid confusion with natural amino acids and degenerate positions). Masked 3-mers were counted and mapped to a 882-dimension column vector, which each dimension representing one of the possible masked 3-mers. Vectors were normalized and scaled by their Z-score.

Once all sequences identified in selection were assembled in column vectors, primary component analysis (PCA) was carried out to identify dimensions (i.e. masked 3-mers) that contributed the most to selection. Sequence reconstruction was carried out for each of the PCA dimensions using positive components above 0.1 (arbitrarily chosen to minimize noise). Reconstruction was carried out by manual inspection assembling selected sequences from the highest to the lowest PCA coefficient. Reconstruction was successful in most cases generating motifs that encompassed both N- and C-terminal arbitrary “**Z**” characters.

Further References

31. Ralser, M. et al. An efficient and economic enhancer mix for PCR. *Biochem. Biophys. Res. Commun.* **347**, 747-751 (2006).