Massively parallel single-amino-acid mutagenesis

Jacob O Kitzman[1,4,5], Lea M Starita[1,5], Russell S Lo[1,2], Stanley Fields[1–3] & Jay Shendure[1]

Random mutagenesis methods only partially cover the mutational space and are constrained by DNA synthesis length limitations. Here we demonstrate programmed allelic series (PALS), a single-volume, site-directed mutagenesis approach using microarray-programmed oligonucleotides. We created libraries including nearly every missense mutation as singleton events for the yeast transcription factor Gal4 (99.9% coverage) and human tumor suppressor p53 (93.5%). PALS-based comprehensive missense mutational scans may aid structure-function studies, protein engineering, and the interpretation of variants identified by clinical sequencing.

ACKNOWLEDGMENTS

We thank P. Brzovic, R. Monnat and members of the Fields and Shendure Labs for helpful discussions. This work was supported by a graduate student research fellowship DGE-0718124 from the US National Science Foundation (to J.O.K.), a US National Institutes of Health (NIH) Pioneer Award #DP1HG007811 (to J.S.) and an NIH Biomedical Technology Research Resource project #P41GM103533 to S.F. S.F. is supported by the Howard Hughes Medical Institute as an investigator.

AUTHOR CONTRIBUTIONS

J.O.K., L.M.S., S.F. and J.S. designed the study and wrote the manuscript. J.O.K., L.M.S. and R.S.L. performed experiments. All authors contributed to and approved the final manuscript.

ONLINE METHODS

Mutagenic primer preparation. Mutagenic primers were electrochemically synthesized on a 12,432-feature programmable DNA microarray and released into solution by CustomArrray[25]. For Gal4 (GI #6325008), codons 2–65 were each replaced with the optimal codon in *Saccharomyces cerevisiae* corresponding to 1 of the 19 other amino acids[26], a stop codon (TAA), or an in-frame deletion, for a total of 1,344 oligos, each synthesized in duplicate (for a total of 2 × 64 × (19 + 1 + 1) = 2,688 array features). For p53 (GI #120407068), codons 1–393 were replaced with fully degenerate bases (“NNN”) during synthesis, such that primer molecules synthesized within a single spot on the array are degenerate for the triplet corresponding to a single residue, for a total of 393 oligos, each synthesized in triplicate (for a total of 3 × 393 = 1,179 array features).

Each primer was designed as a 90-mer, including flanking 15-base adaptor sequences, except for the Gal4 in-frame codon deletion primers, which were designed as 87-mers. Each primer is synthesized sense to the gene, with 33 upstream bases, followed by the codon replacement, and 24 downstream bases. To allow for specific retrieval, a different flanking adaptor pair was used for each subset of mutagenic primers on the array. Gal4 primers were flanked by adaptor sequences “truncL_GAL4DBD” and “truncR_GAL4DBD,” and p53 primers were flanked by “truncL_TP53” and “truncR_TP53” (**Supplementary Table 6**). Mutagenic primer libraries were retrieved by PCR using the respective adaptor pair (“L_TP53”/”R_TP53” or “L_GAL4DBD”/“R_GAL4DBD”), using 10 ng of the starting oligo pool as template using Kapa Hifi Hot Start ReadyMix (“KHF HS RM”, Kapa Biosystems) and following the cycling program “ADO_KHF” (**Supplementary Table 7**). Reactions were monitored by fluorescent signal on a Bio-Rad Mini Opticon real-time thermocycler and were removed after 15 cycles. Amplification products were purified with Zymo Clean & Concentrate 5 columns (Zymo Research). Electrophoresis on a 6% TAE polyacrylamide gel confirmed a single band of ~108 bp for each library, corresponding to the original oligo size plus 18 bp of additional adaptor sequence added by PCR (**Supplementary Fig. 11**).

The resulting oligo pools were further amplified with adaptors modified to contain a deoxyuracil base at the 3′ terminus. This second-round amplification was carried out in 50-µl reactions, using 1 µl of the previous amplification reaction (at a 1:4 dilution in dH_2O) as template, following cycling program “ADO_KR.” Each reaction included 25 µl Kapa Robust Hot Start ReadyMix (which is not inhibited by uracil-containing templates), amplification primers at 500 nM each (“L_”GAL4DBD”/“R_GAL4DBD_U” or “L_TP53”/“R_TP53_U”), and SYBR Green I at 0.5×. Immediately following PCR, each library was denatured at 95 °C for 30 s and then snap cooled on ice. To cleave the “R” adaptors, 2 U USER enzyme mix (New England BioLabs) were added, and each reaction was incubated for 15 min at 37 °C. Finally, each reaction was supplemented by 2.5 µl of a 10 µM stock of the corresponding “L” primer (“L_GAL4DBD” or “L_TP53”), which was followed by one final cycle of annealing/priming/extension. Amplification products were purified as before on Zymo columns. Gel electrophoresis confirmed that each resulting library was a mixture of off-product flanked on both sides by adaptors (108 bp), and the desired product with only “L” adaptors (84 bp; **Supplementary Fig. 11**).

Wild-type template preparation. The full-length Gal4 open reading frame was amplified from genomic DNA of *S. cerevisiae* strain BY4741 and directionally cloned into the yeast shuttle vector p416CYC, a single-copy CEN plasmid with the *CYC1* promoter[27] by digestion with SmaI and ClaI (New England BioLabs), using the InFusion cloning kit (Clontech). Subsequently, an N-terminal truncation was prepared by amplifying residues 1–196 from the original clone using the primer pairs GAL4_CLONE_F and GAL4_NTERM_R and recloning into p416CYC to create p416CYC-Gal4Wt-1-196. This fragment retains the same DNA-binding specificity as full-length Gal4 and is sufficient for transcriptional activation[14]. Enforced expression of full-length Gal4 causes cellular toxicity by aberrantly sequestering the transcriptional machinery in an effect called squelching[28]. A similar effect is observed for Gal4 1–196 (i.e., loss-of-function alleles are more fit than wild-type ones under nonselective growth; **Supplementary Fig. 6**) but to a much lesser degree than for the full-length protein. For p53, a wild-type clone with a C-terminal GFP fusion was purchased from OriGene (#RG200003).

To prepare wild-type sense and antisense strands to serve as templates for mutagenic primer extension, the desired fragments were amplified from plasmid clones by PCR. To select for the sense strand, the reverse primer was phosphorylated to allow for its later degradation by lambda exonuclease, and to select the antisense strand, the forward primer was instead phosphorylated. Furthermore, to minimize undesired carry-through of wild-type copies, in some cases long synthetic tails (38 or 40 nt) were placed on the phosphorylated primer to prevent the resulting 3′ ends of the selected strands from acting as primers during subsequent extension steps. Primers were either ordered with a 5′ phosphate or enzymatically phosphorylated in 10-µl reactions containing 1 µl of 100 µM primer stock, 7 µl H_2O, 1 µl 10× T4 ligase buffer with ATP (NEB), and 10 U T4 polynucleotide kinase (NEB) and incubated for 30 min at 37 °C, followed by heat inactivation for 20 min at 65 °C and 1 min at 95 °C. Wild-type fragments were amplified in 50-µl PCR reactions with forward and phosphorylated reverse primers using Kapa HiFi U+ HotStart Ready Mix (“KHF U+ HS RM”) supplemented with dUTPs to a final concentration of 200 nM. Primers for wild-type template preparation are listed in **Supplementary Table 6**, and amplification used cycling conditions “WT_STRAND_PREP.” For starting template, 200 pg of each wild-type clone plasmid were used. Amplification products were purified by Zymo column, and to select the desired strand, 30 ng of each PCR product were treated for 30 min at 37 °C with 7.5 U lambda exonuclease (NEB) in a 30-µl reaction containing lambda exonuclease buffer at 1× final. Reactions were heat killed for 15 min at 75 °C and purified by Zymo column (5 volumes binding buffer, eluted in 10 µl buffer EB).

Mutagenic primer extension. Next, 2 ng of each primer pool were combined with 3 ng of its respective sense-strand template, raised to 12.5 µl with dH_2O, and mixed with 12.5 µl of KHF U+ HS RM for extension along the dUTP-containing wild-type template by the annealed mutagenic primers. The reaction was subjected to one round of denaturation, annealing, and extension (cycling conditions “PALS_EXTEND”), purified by Zymo column, treated with 1.5 U USER enzyme for 10 min at 37 °C to degrade the wild-type template, and purified again by Zymo column (same conditions).

The resulting strand-extension products were enriched via PCR using the KHF U+ HS RM in 25-µl reactions using the cycling program PALS_AMPLIFY and 3 μl of preceding strand-extension product as template. Reactions were monitored by SYBR Green fluorescence intensity and removed in mid-log phase (13 cycles for Gal4, 10 cycles for p53). The forward and reverse primers corresponding to the sense strand template and the mutagenic adaptor, respectively, were "OUTER_F"/"L_GAL4DBD_U" (for Gal4) or "P53_SENSE_F"/"L_TP53_U" (for p53). An aliquot of each amplification product was visualized by PAGE electrophoresis and appeared as a smear over the expected size ranges (~450–650 bp for Gal4, ~300–1,500 bp for p53; **Supplementary Fig. 11**).

The reverse primer in the preceding amplification step carried a 3′-terminal dUTP, allowing for adaptor excision by treatment with 1 U USER enzyme for 15 min at 37 °C. This reaction was cleaned by Zymo column and eluted in 11.8 μl buffer EB. Next, the respective forward primer was added (0.75 μl at 10 μM) followed by 12.5 μl of KHF HS RM to create sense-strand mutagenized megaprimers with one round of cycling conditions "PALS_EXTEND." For this step, the non-uracil-tolerant PCR mastermix was used to limit amplification of any remaining uracil-containing wild-type strand template. Alternatively, adaptor sequences could be designed to allow excision with Type IIS restriction enzymes.

Sense-strand megaprimers were then purified by Zymo column, annealed to the wild-type antisense strand, and extended to form full-length copies. Each extension reaction contained 3 ng of the sense-stranded megaprimer pool and 1 ng of the wild-type dUTP-containing antisense strand and was performed with KHF U+ HS RM, followed by column cleanup, USER treatment (1.5 U for 10 min at 37 °C), and a second column cleanup, as during the initial mutagenic strand-extension reaction. Finally, the full-length mutagenized copies were enriched by PCR using fully external primers ("OUTER_F"/"GAL4_OUTER_R" or "OUTER_F"/"P53_ANTISENSE_R"), in 25-μl PCR reactions with KHF U+ HS RM with conditions "PALS_AMPLIFY."

PALS library cloning. Gal4 DBD PALS libraries were cloned into p416CYC-bc, a pretagged library of vectors derived from p416CYC in which each clone contains a random 16-mer tag. To prepare p416CYC-bc, a pair of unique restriction sites was placed downstream of the *CYC1* terminator by digesting p416CYC with KpnI-HF (NEB) and inserting a duplex of oligos ("P416CYC_AGEMFE_TOP"/"P416CYC_AGEMFE_BTM") by ligation to create the following series of restriction sites: KpnI-AgeI-MfeI-KpnI. A tag cassette containing a randomized 16-mer ("P416CYC_BC_CAS") was then PCR amplified using primers "P416CYC_AMP_BC_CAS_F"/"P416CYC_AMP_BC_CAS_R" and cycling program "MAKE_BC_CAS" to add priming sites for later tag counting during Gal4 functional selections and to add flanking AgeI and MfeI sites. The resulting tag cassette amplicon was directionally cloned into the modified p416CYC vector by double digestion with AgeI-HF and MfeI-HF (NEB) and transformed into ElectroMax DH10B electrocompetent *Escherichia coli* (Invitrogen), to yield ~9.2 × 10^6 distinctly tagged clones. The resulting library, p416CYC-bc, was expanded by bulk outgrowth and purified by midiprep using the ChargeSwitch Pro Midi kit (Invitrogen). Next, 15 μg of p416CYC-bc were digested with 40 U SmaI (NEB) for 1 h at 25 °C in 60 μl, followed by addition of 20 U ClaI (NEB), digestion for 1 h at 37 °C, and purification by MinElute column (Qiagen). To insert the Gal4 DBD PALS library, 50 ng of the final PALS PCR product were combined with 10 ng SmaI/ClaI linearized p416CYC-bc vector and directionally cloned using the InFusion HD kit (Clontech), as directed. Libraries were transformed by electroporation into 10-beta electrocompetent *E. coli* (NEB), and bulk transformation cultures were expanded overnight in 25 ml LB + ampicillin (50 μg/ml) at 37 °C, shaking at 250 r.p.m. Due to the large number of vector copies present in the cloning reaction, pairing of Gal4 mutant inserts with tag is essentially sampling with replacement; the number of positive clones (~9.0 × 10^5) is less than the number of tags by approximately an order of magnitude, so only ~0.45% of tags are expected to be paired with two different inserts.

Tagged p53 PALS libraries were created in the reverse order: the PALS-mutagenized amplicon was cloned first, and the library was expanded and tags inserted second. The p53 library was cloned into pCMV6-AC-GFP (OriGene) by standard directional cloning in two separate cloning reactions using NotI-HF/BamHI-HF or NotI-HF/KpnI-HF (NEB). Libraries were transformed into 10-beta electrocompetent cells (NEB), combined, expanded overnight, and purified by midiprep as for Gal4. Subsequently, the cloned p53 libraries were linearized at the AgeI site downstream of the hGH poly(A) signal: 2.5 μg of plasmid DNA were digested with 10 U AgeI (NEB) in 50 μl for 1 h at 37 °C and purified by Zymo column. A tag cassette containing a randomized 20-mer was synthesized ("P53_BC_CAS") and PCR amplified for cloning (using primers "P53_AMP_BC_CAS_F"/"P53_AMP_BC_CAS_R"), using KHF RM HS and cycling program "MAKE_BC_CAS." Tags were directionally inserted at the AgeI site by InFusion cloning, as for Gal4, and the resulting plasmid was transformed, expanded in bulk, and purified by midiprep as in the first round of cloning.

Clone subassembly sequencing. To bring the tag cassette into proximity with the mutagenized Gal4 coding sequence (**Supplementary Fig. 10**), 1 μg of the mutant Gal4 plasmid library was digested with 20 U BamHI-HF (NEB) in 1× CutSmart Buffer for 30 min at 37 °C. The digest was cleaned up by Zymo column, and 200 ng of the product were recircularized by intramolecular sticky-end ligation using 1,600 U T4 DNA ligase (NEB) in a 200-μl reaction for 2 h at 20 °C. Following Zymo column cleanup, linear fragments and concatemers were depleted by treatment with 5 U plasmid-safe DNase (Epicentre) for 30 min at 37 °C, and then 30 min at 70 °C. Next, PCR was used to amplify fragments containing the tag cassette at one end, and the mutagenized insert, using 3 μl of the heat-killed recircularization product as template (expected recircularization product and primer pairs shown in **Supplementary Fig. 10a**) and following cycling conditions "PALS_SUBASSEM." Amplification products were purified using Ampure XP beads (1.5× volumes bead/buffer). p53 PALS clone libraries were recircularized following a similar strategy, except that digestions with EcoRI or NotI followed by recircularization were used individually to bring the tag cassette into proximity with the N or C termini, respectively (**Supplementary Fig. 10b**).

To prepare Illumina sequencer-ready subassembly libraries, tag-linked amplicons from the previous step were fragmented and adaptor-ligated using the Nextera v2 library preparation kit (Illumina), with the following modifications to the manufacturer's directions: for each reaction, 1.0 μl Tn5 enzyme "TDE" was combined with 2.0 μl H_2O, 5 μl Buffer 2× TD, and 2 μl of the post-recircularization PCR product. Longer insert sizes were obtained by diluting enzyme TDE up to 1:10 in 1× Buffer TD (a 1:4 dilution was used for the libraries sequenced here). Tagmentation was carried out by incubating for 10 min at 55 °C, followed by library enrichment PCR to add Illumina flow-cell sequences. Libraries were amplified by KHF RM 2× mastermix in 25 µl using a forward primer of NEXV2_AD1 and one of the indexed reverse primers, "SHARED_BC_REV_###." PCR reactions were assembled on ice using as template 2 µl of the transposition reaction (without purification) and cycling omitted the initial strand-displacement step typically used with the Nextera kit (conditions "NEXTERA_SUBASM_PCR"). Last, fixed-position amplicon sequencing libraries starting from the mutagenized insert end of the clone were prepared by adding Illumina flow-cell adaptors directly to the tag-insert amplicons by PCR, using the same PCR conditions but substituting the forward primer "ILMN_P5_SA" for the Nextera-specific forward primer.

Tag-directed clone subassembly. Subassembly libraries were pooled and subjected to paired-end sequencing on Illumina MiSeq and HiSeq instruments, with a long forward read directed into the clone insert (101 bp for HiSeq runs, 325 or 375 bp for MiSeq runs) and a reverse read into the clone tag. Tag-flanking adaptor sequences were trimmed using Cutadapt (obtained from https://code.google.com/p/cutadapt/), and read pairs without recognizable tag-flanking adaptors were excluded from further analysis. Insert-end reads were aligned to the Gal4 or p53 wild-type clone sequence using BWA MEM[29] (with arguments "-z 1 -M"), and alignments were sorted and grouped by their corresponding clone tag. To properly align the programmed in-frame codon deletions included in the Gal4 PALS library, BWA alignments were realigned using a custom implementation of Needleman-Wunsch global alignment with a reduced gap opening penalty at codon start positions (match score = 1, mismatch score = −1, gap open in coding frame = −2, gap open elsewhere = −3, gap extend = −1). A consensus haplotype sequence was determined for each tag-defined read group by incorporating variants present in the group's aligned reads at sufficient depth. Spurious mutations created by sequencing errors, or mutations present at low allele frequency arising from linking two haplotypes to the same tag were flagged and discarded by requiring the major allele at each position (either wild type or mutant) to be present with a frequency of ≥80%, ≥75%, and ≥66%, for read depths ≥20, 10–19, or 3–9, respectively, considering only bases with quality score ≥20. Tag groups with fewer than three reads (Gal4 DBD) or 20 reads (p53) were discarded, as were groups not meeting the major allele frequency threshold across the entire target (Gal4 DBD) or a minimum of 1 kbp (p53). Consensus haplotypes were validated by Sanger sequencing of individual colonies from each tagged plasmid library (**Supplementary Fig. 12** and **Supplementary Table 1**).

Gal4 functional selections. Gal4 DBD PALS libraries were transformed into chemically competent *S. cerevisiae* strain PJ69-4alpha[30] prepared using a modified LiAc-PEG protocol, as previously described[31,32]. After transformation, cells were allowed to recover for 80 min at 30 °C shaking at 250 r.p.m. To select for transformants, cultures were spun down at 2,000*g* for 3 min, resuspended and grown overnight at 30 °C in 40 ml SC medium lacking uracil. Plating 0.25% of the recovery culture before outgrowth indicated a library of ~2×10^5 transformants. Following overnight outgrowth, glycerol stocks were prepared from the transformation culture and stored at −80 °C.

Frozen stocks of yeast carrying the Gal4 DBD PALS library were thawed and recovered overnight in 50 ml SC medium lacking uracil. An aliquot of 1 ml (~1.8×10^6 cells) was pelleted and frozen as the baseline input sample, and equal aliquots were used to inoculate each of four 40-ml cultures of (i) SC medium either lacking uracil (nonselective) or (ii) lacking both uracil and histidine and optionally containing the competitive inhibitor 3-AT (selective; **Supplementary Table 2**). Cultures were maintained at 30 °C and checked at 24 h, 40 h, and 64 h. After reaching log phase (OD_{600} 0.5), each culture was serially passaged by inoculating 1 ml into 40 ml fresh medium.

Input and post-selection cultures were pelleted at 16,000*g* and frozen at −20 °C. Gal4 plasmids were recovered by spheroplast preparation and alkaline lysis miniprep using the Yeast Plasmid Miniprep II kit as directed (Zymo Research). Two-stage PCR was then used to amplify and prepare sequencing libraries to count the plasmid-tagging tags. In the first step, 2.5 µl of miniprep product were used as template in 25-µl reactions with KHF RM HS, with primers flanking the tag cassette ("GAL4_BC_AMP_F"/ "GAL4_BC_AMP_R"), using the program "GAL4_BARCODE_PCR_ROUND1" for 15–17 cycles. The resulting product was used directly as template (1 µl, without cleanup) for the second-stage PCR reaction to add Illumina flow cell–compatible adaptors as well as sample-indexing barcodes to allow pooled sequencing (forward primer "GAL4_ILMN_P5" and reverse primer one of "SHARED_BC_REV_###"). For the second round, the cycling program "GAL4_BARCODE_PCR_ROUND2" was followed for 5–7 cycles. Tag libraries were cleaned up with AmpPure XP beads (2 volumes beads + buffer) and were sequenced across several runs on Illumina MiSeq, GAIIx, and HiSeq instruments (**Supplementary Table 8**), using 25- to 50-bp reads.

Gal4 enrichment scores. Tag reads were demultiplexed to the corresponding sample using a 9-bp index read, allowing for up to two mismatches. Tag reads lacking the proper flanking sequences or containing ambiguous "N" base calls were discarded, and tags were required to exactly match the tag of a single subassembled haplotype. After application of these filters, 18.6% of raw tag reads were discarded. Per-tag histograms were prepared by counting the number of occurrences of each of the remaining tags and normalizing to account for differing coverage over each library by dividing by the sum of tag counts.

We calculated effect scores for each amino acid mutation by summing the read counts of tags corresponding to all the subassembled clones carrying that mutation as a singleton, divided by the equivalent sum for wild-type clones, and taking a log ratio between the selection and input samples:

$$e_{\mathrm{MUT}i} = \log_2\left(\frac{\sum_{\mathrm{TAG}\ j \in \mathrm{MUT}i} r_{\mathrm{SEL},j} + 1}{\sum_{\mathrm{TAG}\ k \in \mathrm{WT}} r_{\mathrm{SEL},k} + 1}\right) - \log_2\left(\frac{\sum_{\mathrm{TAG}\ j \in \mathrm{MUT}i} r_{\mathrm{INPUT},j} + 1}{\sum_{\mathrm{TAG}\ k \in \mathrm{WT}} r_{\mathrm{INPUT},k} + 1}\right)$$

where $r_{\mathrm{SEL},j}$ and $r_{\mathrm{INPUT},j}$ are the read counts of tag j in the selected and input samples, respectively.

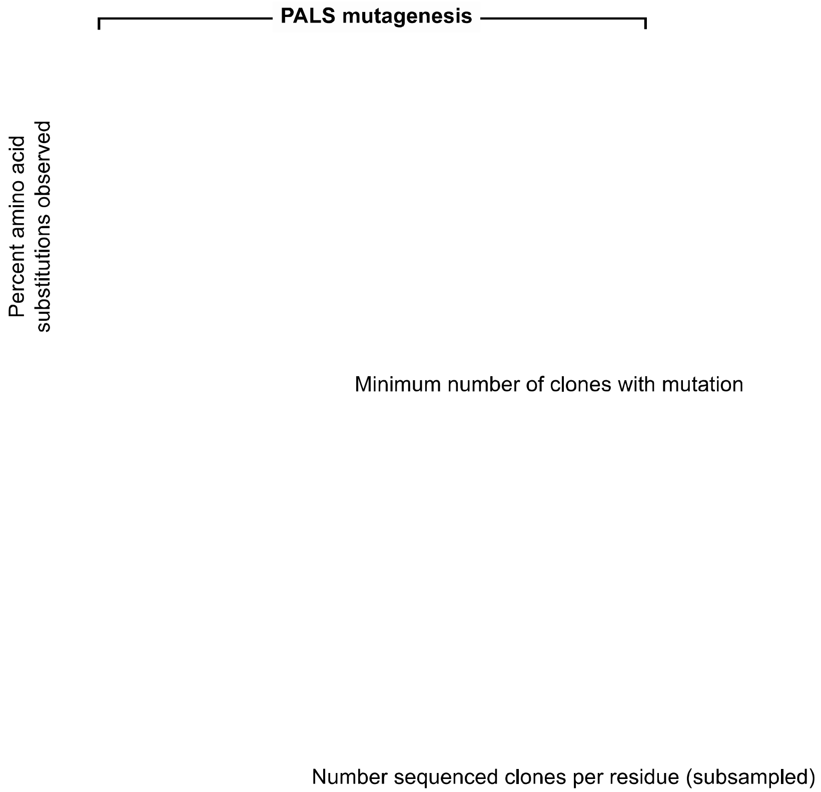
Evolutionarily conserved residues in $\mathrm{Zn_2Cys_6}$ domains were identified by querying HHblits[33] with Gal4 residues 1–70 and were displayed using WebLogo[34]. To compare core and outward-facing residues within the dimerization helix, residues 51–65 were each scored for distance to the overall structure's solvent-exposed surface predicted using MSMS[35] (using the Gal4(1–100) crystal structure, PDB accession 3COQ). Residues with above-median distance to the surface were considered 'core', and those with below-median distance were considered 'exposed', and the $\log_2 E$ values of the two subsets were compared by the Mann-Whitney U-test.

Gal4 effect-size validations. For qualitative validation of effect sizes, eight individual alleles (C14Y, K17E, K25W, K25P, L32P, K43P, K45I, and V57M) were recreated by conventional site-directed mutagenesis and assayed for growth defects by a spotting assay (**Supplementary Fig. 7**). These included loss-of-function (C14Y, K17E, and L32P) and hypomorphic alleles (V57M) from initial screens, which conferred growth rates in the spotting assay that agreed with their relative depletion in the deep mutational scan. We likewise validated a novel predicted hypomorphic allele (K25P) and confirmed the slight growth advantage conferred by three alleles from our bulk measurements (K25W, K43P, and K45I). Each allele was individually introduced into p416CYC-Gal4Wt-1-196 using the Quickchange mutagenesis kit (Agilent) following the manufacturer's directions. Mutant colonies were miniprepped and verified by capillary sequencing and transformed into PJ69-4alpha by LiAc treatment. Following transformation, a single yeast colony transformed by mutant or wild-type Gal4 constructs was picked and expanded in overnight culture and back-diluted to $\mathrm{OD}_{0.2}$ and allowed to return to mid–log phase before spotting tenfold dilutions starting with an equal number of cells onto nonselective plates (SC lacking uracil) or selective plates (SC lacking uracil and histidine, supplemented with 5 mM 3-AT).

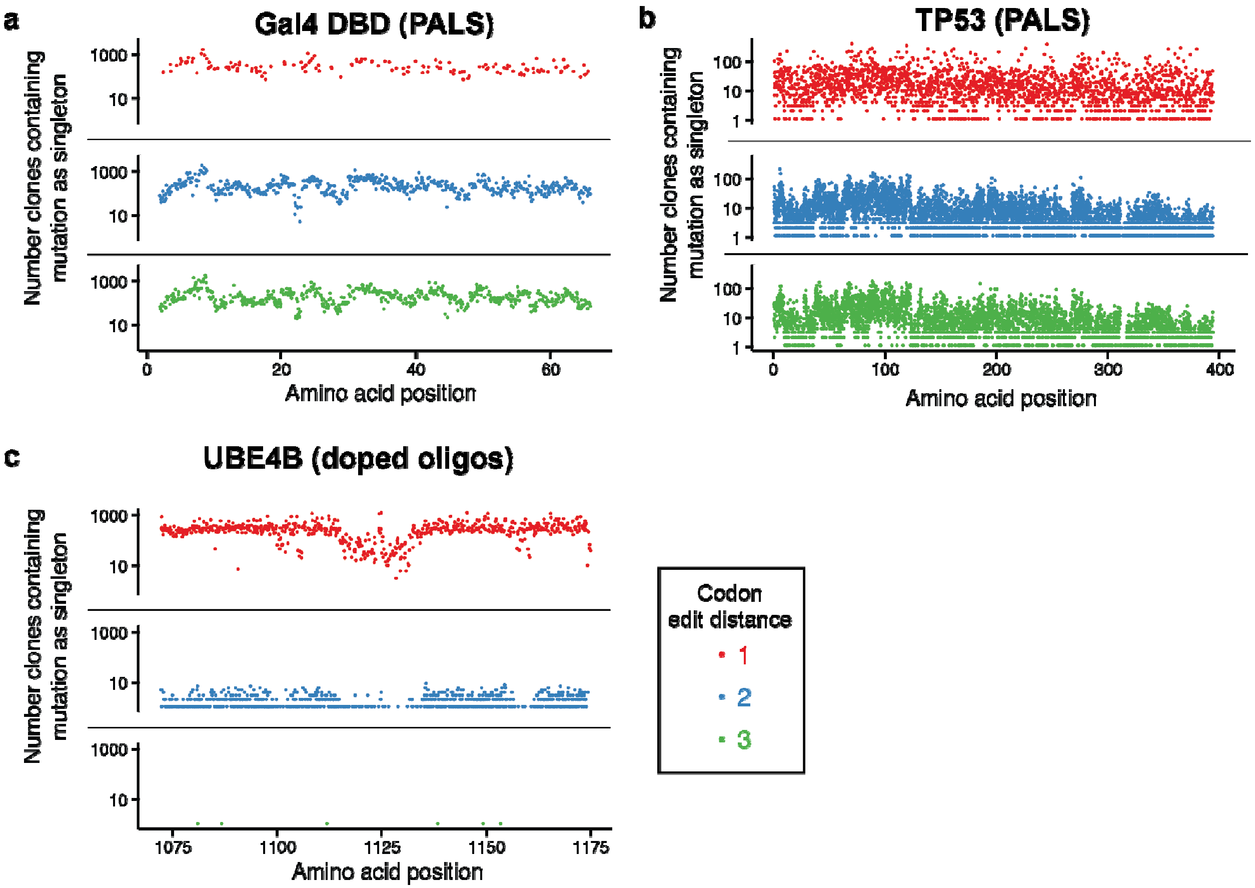
25. Maurer, K. *et al. PLoS ONE* **1**, e34 (2006).
26. Nakamura, Y., Gojobori, T. & Ikemura, T. *Nucleic Acids Res.* **28**, 292 (2000).
27. Mumberg, D., Müller, R. & Funk, M. *Gene* **156**, 119–122 (1995).
28. Gill, G. & Ptashne, M. *Nature* **334**, 721–724 (1988).
29. Li, H. Preprint at http://arxiv.org/abs/1303.3997 (2013).
30. James, P., Halladay, J. & Craig, E.A. *Genetics* **144**, 1425–1436 (1996).
31. Gietz, R.D. & Woods, R.A. *Methods Enzymol.* **350**, 87–96 (2002).
32. Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R. & Fields, S. *RNA* **19**, 1537–1551 (2013).
33. Remmert, M., Biegert, A., Hauser, A. & Söding, J. *Nat. Methods* **9**, 173–175 (2012).
34. Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. *Genome Res.* **14**, 1188–1190 (2004).
35. Sanner, M.F., Olson, A.J. & Spehner, J.C. *Biopolymers* **38**, 305–320 (1996).

with

preparation of mutagenic primers from a DNA microarray. Next, strand extension, strand selection, and PCR

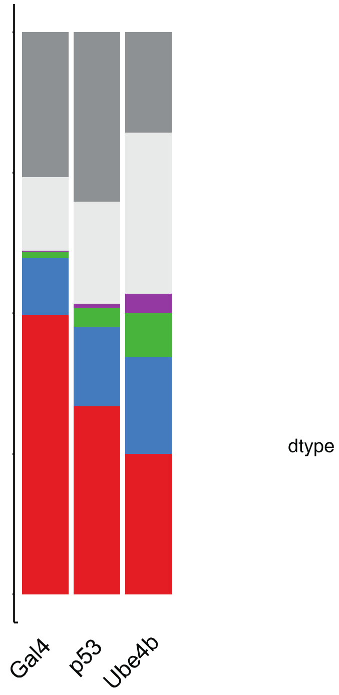


646 clones per residue).



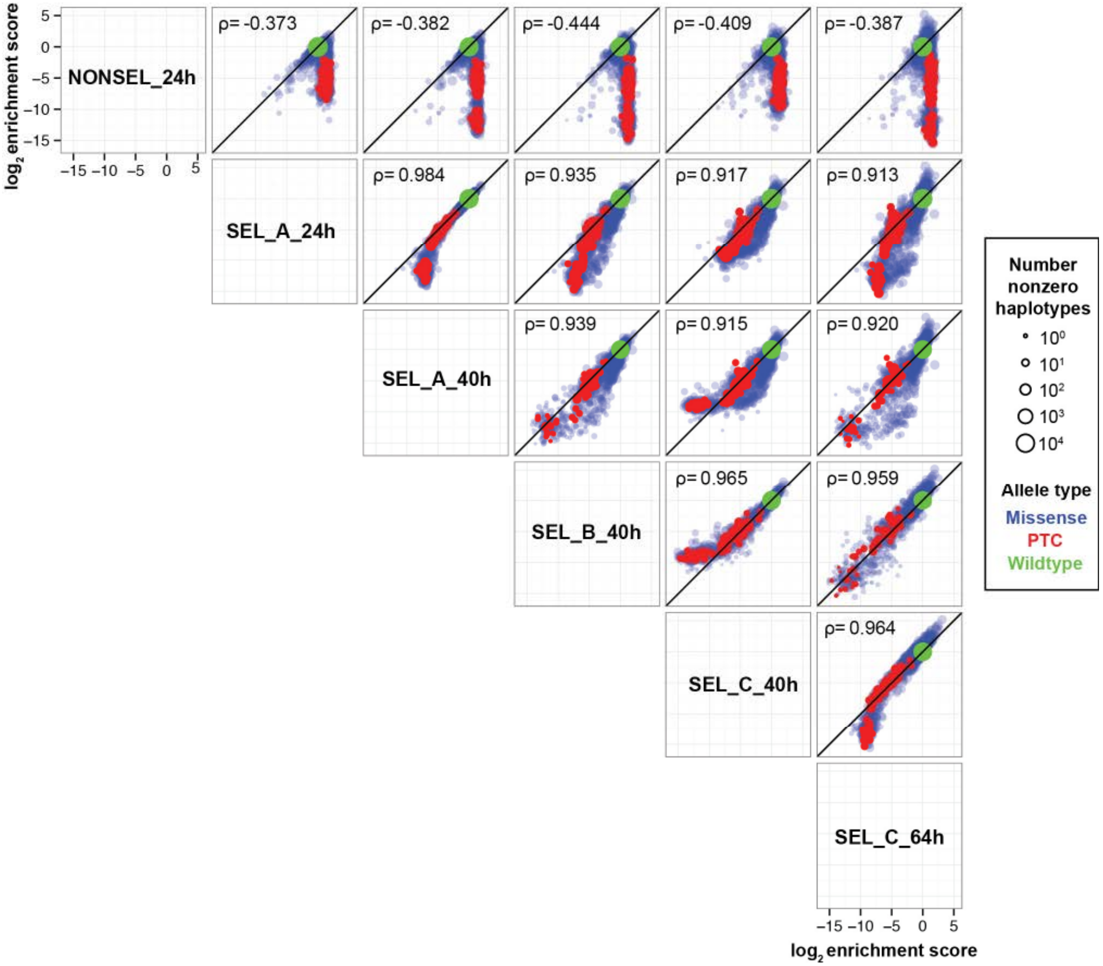
t represents a single codon replacement, shaded by number of base-pair differences, for

or **(a)** Gal4 DBD (n=704,973) and **(b)** p53

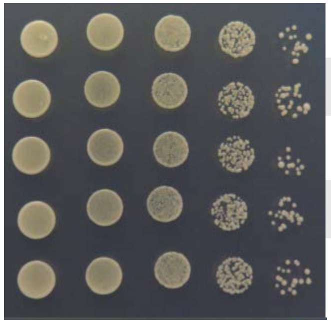


Supplementary Figure 5

Number of coding mutations per clone for PALS (Gal4 and p53) and random mutagenesis (Ube4b) libraries.

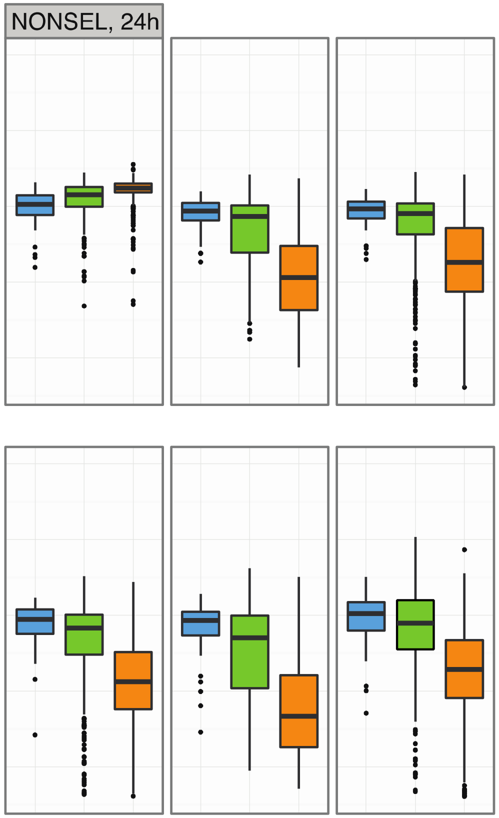


ective conditions, although the magnitude of effect sizes in the former are much larger (90th percentile of absolute log2E values is 11.5 versus 1.66, a >900-fold difference

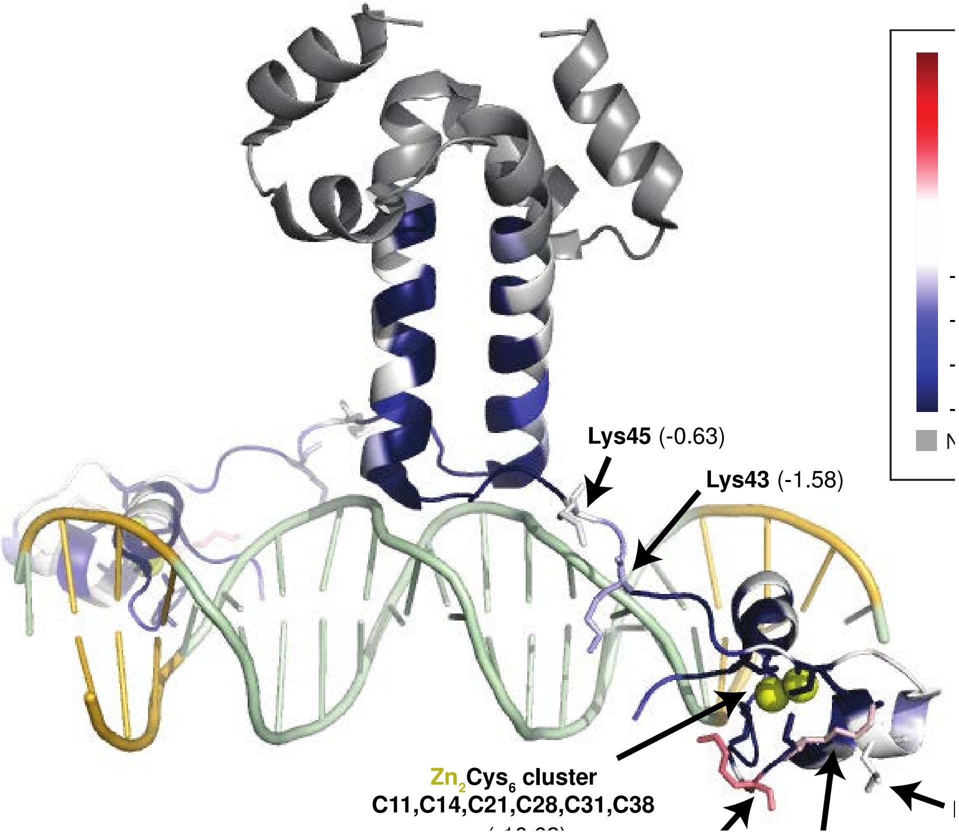


alongside each

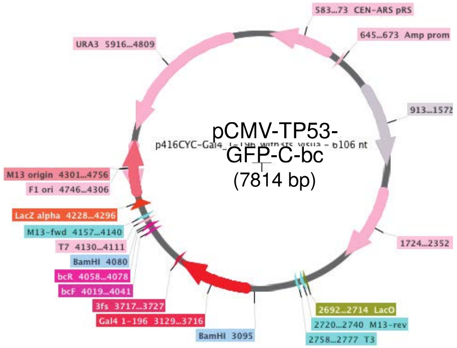
allele (+++, wild-type activity; hypo, hypomorphic; N.D., not determined).



than those that were (group 1).



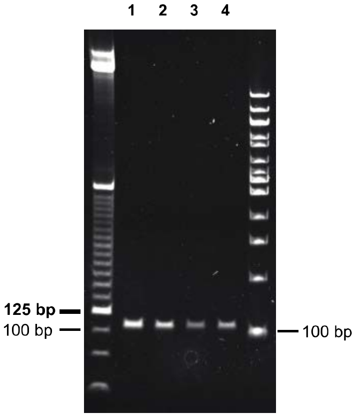
s, are highlighted with median effect size indicated.



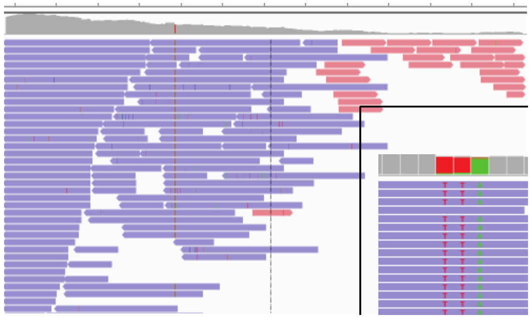
pCMV-TP53-GFP-C-bc
(7814 bp)

s

paired barcode.



s 100bp ladder (NEB).



ne-identifying tag sequence

Supplementary Table 1 | Sanger-sequencing validation of subassembled clones. A total of 40 clones were individually picked and Sanger sequenced across the targeted ORF and associated clone tag, using two reads (Gal4 DBD) or four reads each (p53). Two clones missing from the subassemblies had partially truncated tag sequences (both had single codon replacements with no additional mutations) and one was excluded after failing the allele fraction filter during subassembly. Each of the remaining 37 clone sequences was perfectly concordant with the subassembly consensus sequence bearing the same tag (i.e., no missing or extra mutations). ND, not determined; syn, synonymous mutation.

Barcode from Sanger read	Subassembly and Sanger concordant?	Clone genotype
Gal4 DBD PALS library		
AATGCTGCTGGTGATG	yes	P42D (CCC>GAT)
GTTCTCAGCCGGGCAA	yes	N34G (AAC>GGT), R51K (AGG>AAG)
TGTTAAGGAGACGCGA	yes	T55D (ACA>GAT)
GTAATGAAACTAGGGT	yes	A52X (GCA>TAA)
GAGTAGAGTCGCCGGA	yes	E56G (GAA>GGT)
TAGCATACAAATAAGA	yes	wildtype
AGTGTGGGTGGCATAG	yes	wildtype
ACGTATTAAACAACAC	ND, failed filter	in-frame deletion K18 (programmed)
AATAAGTGACCGGACC	yes	wildtype
TTCATAAAGATCACGT	yes	E8I (GAA>ATT)
TATTTTTAAAAGTGGA	yes	K33D (AAG>GAT)
TCTCAGAGAAATCGTA	yes	S47Y (TCT>TAT), E56L (GAA>TTG)
TAACGTTTTGAATGCG	yes	I7F (ATC>TTT)
p53 PALS library		
GCTTTTGGTACACAGCGTAC	yes	wildtype
ACGTATCGGAAAGCAAATGC	yes	E271S (GAG>TCT), A355F (GCT>TTT), A138syn (GCC>GCT)
CCTGAGTGGGCGACGCCTGA	yes	E2Q (GAG>CAG), Q167syn (CAG>CAA)
AGAAGCTACGTAACAAATTA	yes	wildtype
TCTTGCTTGTGAGGGTGTGG	yes	R202C (CGT>TGT), G245D (GGC>GAC)
ACCCTAAGAGAATACGAGCT	yes	K120L (AAG>TTA)

CTGCGTAGAATGAGCAGGGG	yes	S33F (TCC>TTC), E221L (GAG>TTG)
ATACTCAACATTCTGGACGA	yes	F109R (TTC>CGC), K139syn (AAG>AAA)
GTGCACTCGGGGTAGCAGGG	yes	L137V (CTG>GTC)
TGGTTCCGGACTACAGGAAG	yes	del 1bp frameshift (K371fs)
GCCGCGGGGAGGGCTAGTTA	yes	F212L (TTT>TTA)
CGAGACAATGCAGGTTAGCT	yes	Q165F (CAG>TTT)
TGATATATCGCACCGGAGAA	yes	wildtype
GCACATCCAATACCAGGCGC	yes	E271F (GAG>TTT), K373syn (AAG>AAA)
TTGAGTGGGTCGTGGCAAGA	yes	R175H (CGC>CAC), T81syn (ACA>ACT)
TCCTGACTGCAGGTAGAGGG	yes	G108V (GGT>GTA), G117R (GGG>AGG)
GAACAATGGTACCTGGGAGC	yes	P36S (CCG>TCG), N247F (AAC>TTT)
CCCAAGGTGGGTATAAGGAG	yes	L93Y (CTG>TAC), del 1bp frameshift (S89fs)
GGGAATAAGTAAATGGGCAC	yes	wildtype
GTGGAAAGAGAGGGTAAGAA	yes	A84V (GCC>GTG), A88T (GCC>ACC)
TGGAAGCGCAAAGACTCGAG	yes	P4I (CCG>ATT)
CGAAGGTCGAGTGGTGGACA	yes	del 1bp framehsift (Q191fs), R273C (CGT>TGT), G302syn (GGG>GGA)
AGCTAGGAACGTGAGAAGCC	yes	del 1bp frameshift (I231fs), I232L (ATC>CTC)
TTCTATGCGTGAGTGAGGAC	yes	syn L194:CTT>CTC
GGTATAAAGGGAGCGGGGGC	yes	wildtype
(barcode truncated)	ND, barcode <20 bp	S269P (AGC>CCG)
(barcode truncated)	ND, barcode <20 bp	T253A (ACC>GCT)

Supplementary Table 2 | Gal4 selection cultures and timepoints. SC, synthetic complete

Name	**Media**	**Source**	**Collection timepoint**
INPUT	SC –ura	(Original transformant pool)	0 h
NONSEL_24h	SC –ura	INPUT	24 h
SEL_A_24h	SC –ura –his	INPUT	24 h
SEL_A_40h	SC –ura –his	SEL_A_24h	40 h
SEL_B_40h	SC –ura –his + 0.5 mM 3-AT	INPUT	40 h
SEL_C_40h	SC –ura –his + 1.5 mM 3-AT	INPUT	40 h
SEL_C_64h	SC –ura –his + 1.5 mM 3-AT	SEL_C_40h	64 h

Supplementary Table 3 | Comparison of previously reported activities for Gal4 mutant alleles with effect size measurements in this study. Effect sizes measured in the present study are given as rescaled log2 values (wild-type=0). Jelicic *et al* measured transcriptional activity using a GAL-responsive MEL1 reporter and introduced mutations into a Gal4 fragment containing amino acids 1-100 + 840-881 . Ferdous *et al*[19] performed a similar assay using Gal4 1-147 + 799-1082. Johnston and Dover[15] screened Gal$^-$ mutant alleles within the full-length, native Gal4 locus for activity using a LacZ reporter. ND, not determined.

Allele	Measured activity	log2 effect sizes following selection (rescaled so wt = 0)					
		+his, 24h	-his, 24h	-his, 40h	-his +0.5mM 3AT, 40h	-his +1.5mM 3AT, 40h	-his +1.5mM 3AT, 64h
Jelicic et al[40]							
S5A	~ 70%	0.93	0.27	0.38	-0.42	0.19	1.09
S5D	~ 80%	-0.44	-0.35	-0.39	-0.26	0.12	1.14
S6A	~ wildtype	-1.41	-1.34	-1.56	-1.11	-0.91	-1.65
S6D	~ 75%	0.74	0.38	0.47	0.63	0.21	1.67
S22A	< 5%	0.95	-10.17	-8.90	-9.34	-7.02	-9.14
S22D	~ 25%	-4.91	-6.19	-10.57	-11.00	-11.01	-10.81
S41A	~ 40%	0.70	-0.98	-1.03	-3.42	-4.06	-3.50
S41D	< 5%	1.56	-7.46	-9.94	-11.54	-8.71	-12.35
S47A	~ 25%	0.81	-2.32	-2.69	-9.81	-7.20	-11.61
S59A	~ wildtype	-0.32	-0.85	-0.30	-0.88	-0.62	-0.45
S59D	~ wildtype	-0.81	-0.63	-0.60	-1.21	-0.89	-0.32
Ferdous et al[19]							
S22A	< 5%	0.95	-10.17	-8.90	-9.34	-7.02	-9.14
S22D	~ 20%	-4.91	-6.19	-10.57	-11.00	-11.01	-10.81
K23Q	< 5%	1.21	-2.92	-3.37	-3.94	-3.44	-2.60
K25F	~ wildtype	0.25	0.51	0.62	1.45	2.21	3.57
Johnston and Dover[15]							
C14Y	not detectable	1.30	-3.50	-3.63	-4.34	-3.99	-3.99
R15G	not detectable	1.33	-3.95	-4.56	-7.45	-8.30	-8.37
K17E	not detectable	1.25	-7.44	-10.93	-11.36	-9.56	-12.17
L19P	not detectable	1.47	-3.48	-3.61	-3.20	-2.13	-1.87
S22F	ND (mutant is Gal-)	1.08	-7.81	-12.22	-11.65	-8.75	-12.45
P26L	0.35	0.81	-4.35	-4.78	-10.56	-10.92	-11.95

	(wildtype = 355)						
L32P	not detectable	1.54	-6.17	-6.79	-7.01	-6.59	-5.95
C38G	not detectable	0.45	-2.84	-3.42	-4.12	-4.86	-4.43
S41F	not detectable	1.55	-4.48	-5.07	-6.97	-7.31	-6.65
P42L	not detectable	1.26	-4.14	-4.79	-10.40	-8.37	-10.02
P42S	not detectable	1.05	-2.34	-2.48	-5.63	-6.10	-6.04
S47F	not detectable	1.66	-7.14	-9.99	-11.59	-8.47	-12.40
P48L	ND (mutant is Gal-)	1.21	-6.78	-14.06	-11.50	-8.67	-13.30
P48T	ND (mutant is Gal-)	1.57	-7.22	-10.07	-14.51	-8.81	-13.31
T50I	not detectable	1.51	-4.39	-5.11	-5.81	-6.57	-7.30
R51S	not detectable	1.02	-3.21	-3.56	-2.90	-2.46	-2.22
V57M	1.2 (wildtype = 355)	1.24	-2.07	-2.43	-4.42	-4.63	-4.33

Supplementary Table 4 | Comparison of oligonucleotide synthesis cost, per targeted residue, between PALS and other programmed mutagesis techniques. Cost estimates based upon publicly available list prices for 12k feature 90mer array (CustomArray, Inc.) and 60mer synthesis at the smallest available scale (Integrated DNA Technologies). For both PALS and methods using individually synthesized primers, encoding codon swaps using degenerate NNN triplets required a single oligonucleotide per residue, while specifically programming each codon substitution required 20 (19+1 STOP codon) oligos per residue.

	Each residue replaced by:	
	'NNN' (degenerate, 64 codons)	19 amino acids + STOP
Array-based synthesis (PALS)	\$0.28/residue	\$5.67/residue
Individual column-based synthesis	\$21.00/residue	\$420.00/residue

Supplementary Table 5 | Estimated time required, by step, for PALS mutagenesis library construction. *first two steps can be omitted, to use only a single primer library amplification and cleanup step; QC, quality control checks not depicted on **Supplementary Fig. 1**.

Step	**Hands-on time (min)**	**Total time required (min)**	**Steps from Fig. S1**
Mutagenic primer library amplification I*	15	45	1
Mutagenic primer library cleanup*	5	5	
Mutagenic primer library amplification II	20	50	
Mutagenic primer library cleanup	5	5	
PCR amplify wild-type templates	15	70	2,6
Wild-type template cleanup	5	5	
Wild-type strand selection (lambda digest)	5	50	
Wild-type ssDNA template cleanup	5	5	
Qubit library and template quantification	5	15	QC
Primer library and template gel analysis	10	90	
Mutagenic primer extension on sense template	10	25	3
Wild-type template degradation	5	15	4
Primer extension product cleanup	5	5	
Primer extension product enrichment PCR	15	70	
Post-PCR gel analysis	10	90	QC
Suggested pause point, subtotal	2.25 hr	9.08 hr	
Adaptor cleavage (USER treatment)	5	20	4
PCR product cleanup	5	5	
Forward-strand megaprimer synthesis	5	20	5
PCR product cleanup	5	5	
Megaprimer extension (antisense template)	10	25	7
Wild-type template degradation	5	15	8
Primer extension product cleanup	5	5	
Full length product enrichment PCR	15	70	
PCR product cleanup	5	5	
Subtotal	1.00 hr	2.83 hr	
Total	3.25 hr	11.9 hr	

Supplementary Table 8 | Summary of sequencing performed.

Purpose	Instrument model	# clusters	Read 1 (bp)	Index read (bp)	Read 2 (bp)
Gal4 selection tag counting	Miseq	6,717,738	50	9	NA
	GA IIx	38,979,021	36	9	NA
	Miseq	15,826,768	40	9	NA
	Miseq	15,396,976	40	9	NA
	Hiseq	128,777,071	25	9	NA
	Hiseq	90,831,159	25	9	NA
Subtotal		**296,528,733**			
Gal4 Subassembly	Miseq	8,030,018	325	9	188
	Miseq	21,561,690	325	9	200
	Hiseq	171,094,382	101	9	46
Subtotal		**200,686,090**			
TP53 Subassembly	Miseq	4,703,001	325	9	185
	Miseq	8,969,328	325	9	185
	Hiseq	90,525,981	101	9	101
	Miseq	23,788,171	375	9	104
Subtotal		**127,986,481**			

Supplementary Note 1 | Multiple-mutation analysis

Subassembled clones with multiple mutations examined to investigate the underlying cause of the secondary mutations. For the Gal4 DBD PALS library, these were dominated by PCR chimeras (52% among clones with secondary mutations) and synthesis errors (24%), as estimated by counting clones bearing two programmed mutations, or one programmed mutation and secondary mutations within the boundaries of the corresponding mutagenic primer. Chimerism is a technical challenge commonly encountered while amplifying libraries of homologous sequences[37], when incomplete strand extension products in one cycle of amplification act as primers in the subsequent cycle. Future optimization efforts will be directed at quantifying and mitigating this phenomenon by manipulating input template concentration and minimizing amplification cycles, or alternatively using droplet PCR[38]. To reduce the impact of synthesis errors, PALS uses short oligonucleotides (90 nt), but it will nevertheless benefit from ongoing developments in high-fidelity synthesis[39]. In addition, as single-base deletions are the dominant synthesis error mode[40], stringently size-selecting primer libraries may further enrich for primers lacking undesirable secondary mutations. Another strategy would fuse libraries in-frame to a selectable marker in the bacterial cloning host, although our preliminary observations suggest that such selection is inefficient for proteins that do not fold or express well in E. coli. For p53, because codon substitutions were encoded as “NNN”, the origin of secondary mutations could not be distinguished between synthesis errors, PCR errors, or chimerism between fragments each bearing a single codon swaps.

Although PALS is intended to create single-mutant clones, for applications such as protein engineering, it may be useful to obtain multiple mutations per copy. This could be accommodated by applying PALS serially, to first create a library of single-mutant copies which would then be used as the starting template for the second round. In the context of typical-length genes, the multi-mutation space is so large (e.g., for TP53 double mutants, choose(393,2)*19*19=2.78x10^7

possibilities) that it may be technically impractical to construct, much less survey, the entire space. By serially applying PALS, however, it could be possible to focus on a defined subspace using a subset of mutagenic primers in either or both rounds.