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.

Site-directed mutagenesis is an indispensable tool for sequence-structure-function studies[1]. However, conventional approaches such as Kunkel mutagenesis and its refinements[2] traditionally target only one site at a time. Consequently, many separate reactions are required to systematically mutagenize a protein sequence for subsequent functional analysis by alanine scanning[3] or more recent massively parallel methods.

One such method, deep mutational scanning[4], subjects large libraries of mutants to assays that select for the function of the protein. Digital counting via deep sequencing of libraries before and after functional selection is used to quantify the enrichment or depletion of individual mutants as a proxy for functional impact. These approaches typically build mutant libraries via doped oligonucleotide synthesis[4,5], in which the targeted region is synthesized with a tunable error rate. However, frameshifting deletion errors limit the length of sequence that can be directly synthesized. Error-prone PCR represents an alternative, but it requires empirical tuning to reach a desired mutational load and suffers from bias[6]. A shared limitation of these methods is that only a minority of the codon mutational space can be accessed through single-base mutations (for example, 31% for p53).

Scalable methods for programmed mutagenesis are needed in order to enable deep mutational scans of longer sequences[7–9]. Recent advances[10–12] provide a degree of multiplexing to this end but remain laborious and cost-prohibitive, as they require individual synthesis of mutagenic primers or are limited in their scope by targeting only a few residues at a time, necessitating serial tiling over the target.

To overcome these limitations, we developed PALS, which combines low-cost, microarray-based DNA synthesis with overlap-extension mutagenesis to introduce one and only one mutation per cDNA template in a massively parallel fashion. The PALS workflow begins with on-array synthesis of mutagenic primers tiling a target, with each bearing a mutation (e.g., codon swap) near its center (step 1; **Fig. 1a**). Each primer library is designed with flanking adaptors, allowing specific subsets to be retrieved by PCR. Downstream adaptors are removed (**Supplementary Fig. 1**), and pools of tailed primers are annealed and extended along a linear wild-type sense strand marked by deoxyuracil (dU; step 2), which is then degraded with uracil-DNA-glycosylase (UDG) and exonuclease VIII. The nested strand-extension product is PCR amplified using an upstream forward primer and a reverse primer corresponding to the adaptor sequence at the 5′ end of each mutagenic primer (step 3). After amplification, the adaptor sequence is clipped, and the resulting mutagenized megaprimer is extended to full length along a wild-type antisense strand (step 4). Residual wild-type strands are again degraded with UDG, and the full-length library of mutant cDNAs is enriched by PCR (step 5) and cloned.

Assessing the rates of programmed and off-target mutagenesis requires that the resulting library be sequenced. Deep shotgun sequencing may detect all programmed mutations, but because currently available sequencing reads are short, multiple mutations on the same clone cannot be phased. Consequently, a neutral substitution could be wrongly counted as highly deleterious when coupled to a nonsense mutation elsewhere on the same clone. To obtain full-length sequences for PALS-mutagenized clones, we used 'subassembly'[13], in which each mutant cDNA clone in a complex library is individually coupled with a random molecular 'tag' (**Fig. 1b**). Paired-end reads are obtained with a fixed end reporting the tag sequence and a shotgun end derived randomly from the insert. Shotgun reads are then grouped by tag to yield an accurate full-length consensus haplotype that is longer than the constituent reads and corrects random sequencing errors (37/37 clones validated by Sanger; **Supplementary Table 1**). After haplotype-resolved sequencing of the mutant clone pool, molecular tags may be counted in bulk to quantify allelic enrichment or depletion following function-dependent selection, thus obviating deep sequencing of the longer clone inserts after each selection step.

As a proof of concept, we constructed a PALS library for the DNA-binding domain (DBD) of Gal4, a yeast transcription factor.

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. [3]Department of Medicine, University of Washington, Seattle, Washington, USA. [4]Present address: Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to J.O.K. (kitzmanj@umich.edu) or J.S. (shendure@uw.edu).

RECEIVED 24 JULY 2014; ACCEPTED 18 NOVEMBER 2014; PUBLISHED ONLINE 5 JANUARY 2015; CORRECTED AFTER PRINT 10 APRIL 2017; DOI:10.1038/NMETH.3223

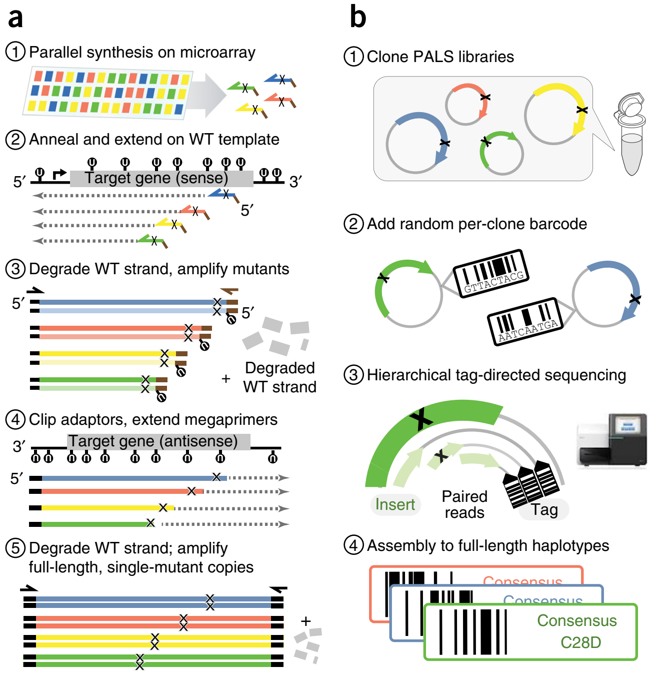
a
① Parallel synthesis on microarray
② Anneal and extend on WT template
5′
Target gene (sense)
3′
5′
③ Degrade WT strand, amplify mutants
5′
5′
Degraded
WT strand
+
④ Clip adaptors, extend megaprimers
3′
Target gene (antisense)
5′
⑤ Degrade WT strand; amplify
full-length, single-mutant copies
+
b
① Clone PALS libraries
② Add random per-clone barcode
GTTACTACG
AATCAATGA
③ Hierarchical tag-directed sequencing
Insert
Paired
reads
Tag
④ Assembly to full-length haplotypes
Consensus
Consensus
Consensus
C28D

Figure 1 | Programmed allelic series (PALS) mutagenesis in a single volume reaction. (**a**) PALS mutagenesis. Primers are synthesized in parallel on a microarray, tiling a target sequence of interest and bearing programmed mutations (“X”), for example, to make specific or random codon substitutions or tiling deletions. Programmed mutations are introduced by primer extension on a degradable wild-type (WT) template (marked with deoxyuracil, “U”) followed by PCR amplification with primers directed to the gene flanks (black) or to adaptor sequences within the mutagenized strands (brown). A final PCR step yields full-length copies incorporating a single programmed mutation per copy. (**b**) Library sequencing and tagging. Mutant libraries are cloned, with each clone receiving a unique molecular tag sequence. The library is subjected to hierarchical shotgun sequencing, with paired-end reads interrogating the target gene insert from one end and the molecular tag from the other, to yield a set of consensus haplotypes and associated tags.

We targeted each Gal4 DBD codon (residues 2–65) for replacement either by the yeast-optimized codon for each of the 19 other amino acids or by a premature termination codon. After cloning and subassembly, ~47% of full-length haplotypes carried one and only one programmed mutation on an otherwise wild-type background (**Table 1**). Among these ‘clean’ clones, 99.9% (n = 1,342) of programmed single-codon replacements were observed at least once, and 99.7% were observed at least five times (**Supplementary Fig. 2**). We also programmed in-frame deletions of each codon, all of which we observed in the resulting library.

To assess PALS’ scalability from a single domain to a full-length cDNA, we next targeted the entire coding sequence of human p53. In contrast to Gal4, for which we explicitly specified each mutant codon, we targeted p53 codons for replacement by degenerate (‘NNN’) triplets, reducing the microarray features required to the number of codons (393 for p53) and allowing access to synonymous variants. We observed a lower rate of sequence-verified single-mutant haplotypes (33%, n = 216,714) owing to the greater potential for secondary errors on longer templates, largely due to PCR chimerism (**Supplementary Note**). Despite the reduced purity and lower sequencing depth relative to the Gal4 library, we still observed 7,345 of 7,860 (93.4%) possible amino acid substitutions in p53 as clean, single-mutant clones.

Mutational coverage by PALS was relatively uniform with a moderate bias toward the N terminus (1.1-fold for Gal4 DBD and 2.2-fold for p53; **Supplementary Fig. 3**). For comparison, we reanalyzed a random mutant library[5] constructed by doped synthesis. That library comprised 1.12 million clones, of which 25.0% contained a single-codon mutation. Codon substitutions requiring 2- or 3-bp changes, well represented within PALS libraries, were rare or absent in the randomized library (**Supplementary Fig. 2**). Simulations indicate that varying the randomized mutagenesis rate would partially restore coverage of these substitutions, at the cost of creating many more clones with multiple mutations including nonsense codons (**Supplementary Fig. 4**). PALS libraries also had fewer insertion or deletion (indel)-bearing clones (13.2–18.2% versus 28.6% for the randomized library; **Supplementary Fig. 5**), most of which encode frameshifts that are uninformative for functional analysis.

We next used PALS to perform a comprehensive deep mutational scan. We introduced the Gal4 DBD PALS library (fused to an additional 131-amino-acid (aa) wild-type fragment sufficient for transcriptional activation[14]) into a two-hybrid reporter strain in which *GAL4* is deleted and the *HIS3* gene is under the control of the *GAL1* promoter. Thus, growth on medium lacking histidine was conditional upon the ability of the introduced Gal4 DBD mutant to bind to and activate *HIS3* expression. We modulated selection stringency by addition of 3-amino-1,2,4-triazole (3-AT), a competitive inhibitor of His3. After selection for Gal4 function, we performed deep sequencing of the linked tags to quantify the enrichment or depletion of each Gal4 mutant.

We collected 296.5 million tag reads across the input library and six selection time points (**Supplementary Table 2**). We summed tag counts across clones bearing the same single-amino-acid mutation and calculated per-mutation effect sizes ($\log_2 E$) for the 98.2% of mutations (1,320/1,344) that were each represented by at least four distinct tagged clones in the nonselected library. After two rounds of yeast outgrowth under stringent conditions (t = 64 h in histidine-free medium supplemented with 1.5 mM 3-AT), the enrichment score distribution was shifted downward, with 57.3% of single-amino-acid mutants strongly depleted ($\log_2 E < -3$). As expected, premature stop mutations were nearly uniformly deleterious under selective conditions but not permissive conditions (median $\log_2 E = -5.75$ and $+1.33$, respectively). About one-third of the residues (19–27 of 64, depending on selection time point) were strongly intolerant to mutation, having a median effect size for nontruncation mutants at least as low as the overall median of premature truncation mutants. Per-mutation effect sizes were well-correlated across time points and replicates (Spearman’s ρ = 0.917–0.984; **Supplementary Fig. 6**) and were validated well

Table 1 | Summary of sequence-verified haplotypes by mutation status

	Gal4 DBD clones (%)	p53 clones (%)
Designed (single coding mutation)	328,871 (47)	216,714 (33)
Designed plus secondary mutation	149,311 (21)	227,592 (35)
Wild-type	171,475 (24)	195,000 (30)
Only nonprogrammed mutations[a]	55,316 (8)	7,633 (1)
Total # of sequence-verified haplotypes	704,973	646,939

[a] A point or indel mutation observed in clones but not programmed in mutagenic primers.

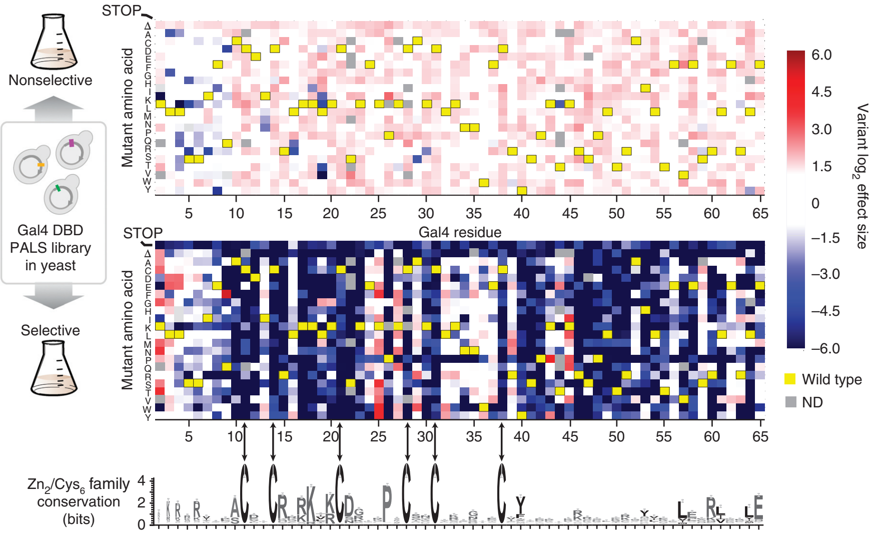

Figure 2 | *En masse* functional selection of Gal4 DBD PALS library highlights residues and mutations critical for transcriptional activity. Top, sequence-function map of mutation effect sizes across Gal4 DBD residues 2–65 for all programmed amino acid substitutions, following outgrowth without selection (SC medium –uracil, after 24 h). Center, sequence-function map under stringent selection for Gal4 (SC –uracil –histidine +1.5 mM 3-AT, after 64 h). STOP, premature stop codon; Δ, in-frame codon deletion. Sequence-function maps are shaded by the $\log_2$ effect size for each residue and substitution, ranging from improved growth over wild type (red), to equivalent to wild type (white), to slower growth than wild type (blue). Yellow and gray boxes denote the wild-type residue or insufficient data, respectively (minimum four distinct tagged haplotypes per codon substitution required in the nonselected library). Bottom, evolutionary conservation among Zn_2/Cys_6 family members (plotted in bits) confirms selective constraint to maintain the six domain-defining cysteines (indicated by arrows).

by qualitative spotting assays (**Supplementary Fig. 7**) and by agreement with previous reports (**Supplementary Table 3**).

The resulting profile of functional constraint (**Fig. 2** and **Supplementary Data**) encompasses loss-of-function alleles from initial genetic screens[15] and key features from structural studies[16]. Gal4 binds DNA as a homodimer via a Zn_2Cys_6-class domain centered on a pair of Zn^{2+} ions, which help to maintain the fold of the DNA-binding residues. Substitution at any of six chelating cysteines completely disrupted function, a result consistent with their essential role and strong conservation. More broadly, other conserved residues were significantly less tolerant to substitution during selective outgrowth ($P < 1.6 \times 10^{-7}$ comparing per-residue mean $\log_2 E$, Mann-Whitney U; **Supplementary Fig. 8**).

Superimposed on the crystal structure[17] (residues 1–100; **Supplementary Fig. 9**), these data suggest additional key molecular interactions. As expected, core residues within the dimerization helix were less mutation tolerant than outward-facing ones ($P < 1.6 \times 10^{-4}$, Mann-Whitney U). In the unstructured linker (residues 41–50), a bend at Pro48 aids in positioning the dimerization helix over the DNA minor groove[16]. Either of two nearby lysine residues (Lys43 and Lys45) could be mutated to proline without deleterious effects (**Supplementary Fig. 7**). Except in the disordered N terminus, proline substitutions were highly deleterious. For instance, leucine 32 is central to one of the two metal-binding domain α-helices and showed little constraint (mean $\log_2 E = -0.04$), aside from replacement with proline, which completely abrogates Gal4 DNA binding[15].

This trend is broadly observed in deep mutational scans of other proteins, likely reflecting disruption of protein secondary structure due to the proline residue kinking the backbone[18]. Within the Gal4 DBD linker region, however, additional prolines may be beneficial by decreasing the flexibility between the dimerization and zinc-containing regions, making DNA binding and transcriptional activation more entropically favorable. Similar to most proline mutations, in-frame codon deletions were generally deleterious, with the notable exceptions of Lys25 and Lys27, both outward-facing lysines located near proposed sites of post-translational modification in the loop between metal-binding domain helices[19]. Proline mutations or in-frame deletions that are disruptive at otherwise mutation-tolerant residues (for example, 32–37) can thus serve to distinguish residues that are structurally important but do not participate in catalysis or critical post-translational modifications. Although such mutations are unlikely to arise naturally, their inclusion may nevertheless provide valuable insight.

PALS enables near-comprehensive, single-amino-acid mutagenesis of a protein-coding sequence in a single reaction volume within 2 d, and its use of microarray synthesis markedly reduces reagent costs (**Supplementary Tables 4** and **5**). Other functional screens exploiting programmed oligonucleotide libraries[20,21] have been limited to shorter sequence elements owing to synthesis length constraints (100–200 nt), which PALS overcomes by highly multiplexed overlap-extension PCR on a wild-type template. Analysis of long PALS targets is presently limited by constraints on subassembly, but there may be workarounds (**Supplementary Fig. 10**).

Genome-editing technologies such as CRISPR-Cas have recently enabled large-scale knockout screens[22,23] and saturation mutagenesis of short exons[24] at their native genomic loci. Future applications of these editing approaches, using PALS-mutagenized copies as a homology-directed repair template pool, may enable the systematic analysis of genomic mutations across human coding genes. The combination of PALS mutagenesis, functional selection and deep sequencing provides a general framework to dissect the allelic heterogeneity of human genes and a path toward ‘precomputed’ functional annotation of the growing catalogs of variants of unknown significance.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Sequence Read Archive: PRJNA268398. Addgene: p53 allelic series library, 61040.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

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.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Botstein, D. & Shortle, D. *Science* **229**, 1193–1201 (1985).
2. Kunkel, T.A. *Proc. Natl. Acad. Sci. USA* **82**, 488–492 (1985).
3. Cunningham, B.C. & Wells, J.A. *Science* **244**, 1081–1085 (1989).
4. Fowler, D.M. & Fields, S. *Nat. Methods* **11**, 801–807 (2014).
5. Starita, L.M. *et al. Proc. Natl. Acad. Sci. USA* **110**, E1263–E1272 (2013).
6. Wong, T.S., Roccatano, D., Zacharias, M. & Schwaneberg, U. *J. Mol. Biol.* **355**, 858–871 (2006).
7. Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D. & Bolon, D.N.A. *J. Mol. Biol.* **425**, 1363–1377 (2013).
8. Qi, H. *et al. PLoS Pathog.* **10**, e1004064 (2014).
9. Firnberg, E., Labonte, J.W., Gray, J.J. & Ostermeier, M. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
10. Firnberg, E. & Ostermeier, M. *PLoS ONE* **7**, e52031 (2012).
11. Hietpas, R.T., Jensen, J.D. & Bolon, D.N.A. *Proc. Natl. Acad. Sci. USA* **108**, 7896–7901 (2011).
12. Jain, P.C. & Varadarajan, R. *Anal. Biochem.* **449**, 90–98 (2014).
13. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. *Nat. Methods* **7**, 119–122 (2010).
14. Ma, J. & Ptashne, M. *Cell* **48**, 847–853 (1987).
15. Johnston, M. & Dover, J. *Genetics* **120**, 63–74 (1988).
16. Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S.C. *Nature* **356**, 408–414 (1992).
17. Hong, M. *et al. Structure* **16**, 1019–1026 (2008).
18. Chou, P.Y. & Fasman, G.D. *Annu. Rev. Biochem.* **47**, 251–276 (1978).
19. Ferdous, A. *et al. Mol. Biosyst.* **4**, 1116–1125 (2008).
20. Patwardhan, R.P. *et al. Nat. Biotechnol.* **27**, 1173–1175 (2009).
21. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T.S. *Nucleic Acids Res.* **42**, e112 (2014).
22. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. *Science* **343**, 80–84 (2014).
23. Shalem, O. *et al. Science* **343**, 84–87 (2014).
24. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C. & Shendure, J. *Nature* **513**, 120–123 (2014).

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$_2$O) 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$_2$O, 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 µm. 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$_2$O, 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:

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

doi:10.1038/nmeth.3223

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 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 minipreppped 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 $OD_{0.2}$ 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).

doi:10.1038/nmeth.3223

Corrigendum: Massively parallel single-amino-acid mutagenesis

Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields & Jay Shendure

***Nat. Methods* 12, 203–206 (2015); published online 5 January 2015; corrected after print 10 April 2017**

In the version of this article initially published, the unit (nM) for the dUTP concentration described in the "Wild-type template preparation" section of the Online Methods was incorrect. The correct unit should be μM. The error has been corrected in the HTML and PDF versions of the article as of 10 April 2017.