Massively Parallel Single Amino Acid Mutagenesis

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

[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

Abstract

Random mutagenesis methods only partially cover the mutational space, and are constrained by DNA synthesis length limitations. Here, we demonstrate 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 indispensible tool for sequence-structure-function studies[1]. However, conventional approaches like 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, frame-shifting deletion errors limit the length of sequence that can be directly synthesized. Error-prone PCR represents an alternative, but 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 (*e.g.*, 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 ("programmed allelic series"), 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 (Fig. 1a, step 1). 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). The remaining 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 UDG-degraded, 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 "sub-assembly"[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, 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. 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 STOP. 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%) of the desired amino acid substitutions in p53 as clean, single-mutant clones.

Mutational coverage by PALS was relatively uniform with a moderate bias towards the N-terminus (1.1-fold for Gal4 DBD; 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-bp 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 indel-bearing clones (13.2-18.2% versus 28.6% for the randomized library, Supplementary Fig. 5), most of which encode frame-shifts 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-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 media 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 timepoints (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 (1320/1344) that were each represented by at least four distinct tagged clones in the non-selected library. After two rounds of yeast outgrowth under stringent conditions (t=64 hours in –histidine media 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 non-truncation 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 validated well by qualitative spotting assays (Supplementary Fig. 7) and by agreement with previous reports (Supplementary Table 3).

The resulting profile of functional constraint (Fig. 2, Supplementary Dataset 1) 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, consistent with their essential role and strong conservation. More broadly, other conserved residues were significantly less tolerant to substitution during selective outgrowth ($P<1.6\times10^{-7}$ 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 ($P<1.6\times10^{-4}$, Mann-Whitney U). In the unstructured linker (residues 41-50), a bend at proline 48 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 alpha 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 (e.g., 32-37) can thus serve to distinguish residues that are structurally important but 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 two days, while 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 due 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 "pre-computed" functional annotation of the growing catalogs of variants of unknown significance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References Cited

1. Botstein D, Shortle D. Science. 1985; 229:1193–1201. [PubMed: 2994214]

2. Kunkel TA. Proc Natl Acad Sci USA. 1985; 82:488–492. [PubMed: 3881765]

3. Cunningham BC, Wells JA. Science. 1989; 244:1081–1085. [PubMed: 2471267]

4. Fowler DM, Fields S. Nature Methods. 2014; 11:801–807. [PubMed: 25075907]

5. Starita LM, et al. Proceedings of the National Academy of Sciences. 2013; 110:E1263–72.

6. Wong TS, Roccatano D, Zacharias M, Schwaneberg U. J Mol Biol. 2006; 355:858–871. [PubMed: 16325201]

7. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. J Mol Biol. 2013; 425:1363–1377. [PubMed: 23376099]

8. Qi H, et al. PLoS Pathog. 2014; 10:e1004064. [PubMed: 24722365]

9. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. Molecular Biology and Evolution. 2014; 31:1581–92. [PubMed: 24567513]

10. Firnberg E, Ostermeier M. PLoS ONE. 2012; 7:e52031. [PubMed: 23284860]

11. Hietpas RT, Jensen JD, Bolon DNA. Proceedings of the National Academy of Sciences. 2011; 108:7896–7901.

12. Jain PC, Varadarajan R. Analytical Biochemistry. 2014; 449:90–98. [PubMed: 24333246]

13. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Nature Methods. 2010; 7:119–122. [PubMed: 20081835]

14. Ma J, Ptashne M. Cell. 1987; 48:847–853. [PubMed: 3028647]

15. Johnston M, Dover J. Genetics. 1988; 120:63–74. [PubMed: 3065140]

16. Marmorstein R, Carey M, Ptashne M, Harrison SC. Nature. 1992; 356:408–414. [PubMed: 1557122]

17. Hong M, et al. Structure. 2008; 16:1019–1026. [PubMed: 18611375]

18. Chou PY, Fasman GD. Annu Rev Biochem. 1978; 47:251–276. [PubMed: 354496]

19. Ferdous A, et al. Mol BioSyst. 2008; 4:1116. [PubMed: 18931787]

20. Patwardhan RP, et al. Nat Biotechnol. 2009; 27:1173–1175. [PubMed: 19915551]

21. Melnikov A, et al. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]

22. Wang T, Wei JJ, Sabatini DM, Lander ES. Science. 2014; 343:80–84. [PubMed: 24336569]

23. Shalem O, et al. Science. 2014; 343:84–87. [PubMed: 24336571]

24. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Nature. 2014; 513:120–123. [PubMed: 25141179]

Figure 1.
Programmed Allelic Series (PALS) mutagenesis in a single volume reaction. **(a)** Primers are synthesized in parallel on a microarray, tiling a target sequence of interest and bearing programmed mutations ("X"), e.g., to make specific or random codon substitutions or tiling deletions. Programmed mutations are introduced by primer extension on a degradable wild-type 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)** 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.
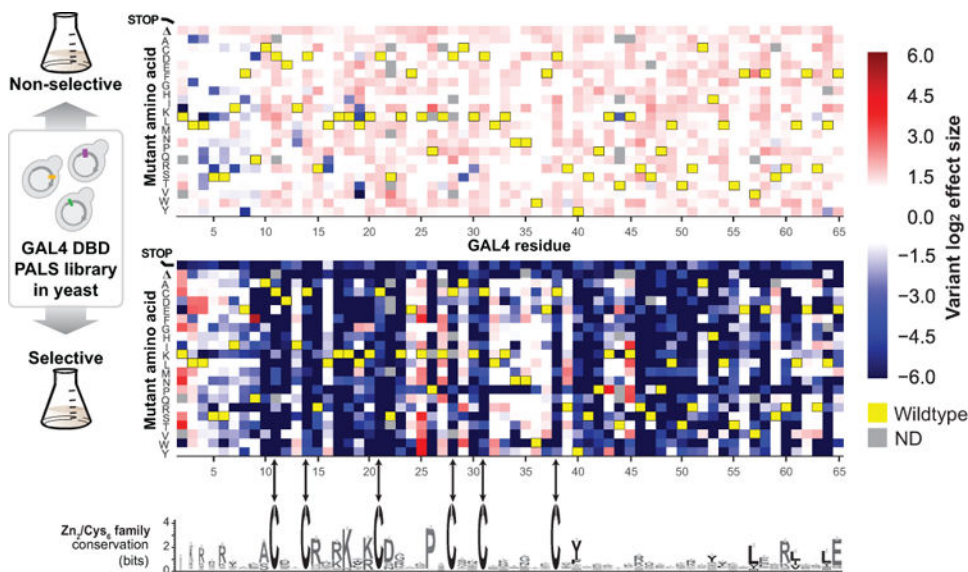
Figure 2.
En masse functional selection of Gal4 DBD PALS library highlights residues and mutations critical for transcriptional activity. Sequence-function maps of mutation effect sizes across Gal4 DBD residues 2-65 (rows) for all programmed amino acid substitutions (columns; STOP: premature stop codon, : in-frame codon deletion) following outgrowth either without selection (upper: SC – uracil, after 24 h) or under stringent selection for Gal4 (lower: SC – uracil – histidine + 1.5 mM 3-AT, after 64 h). Sequence-function maps are shaded by the log2-effect size for each residue and substitution, ranging from improved growth versus wild-type (red), 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 non-selective library). Below, evolutionary conservation among Zn2/Cys6 family members (plotted in bits), confirms selective constraint to maintain the six domain-defining cysteines (indicated by arrows).

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 non-programmed mutations[*]	55,316 (8%)	7,633 (1%)
Total # sequence-verified haplotypes	**704,973**	**646,939**

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