Published in final edited form as:
Trends Biotechnol. 2011 September ; 29(9): 435–442. doi:10.1016/j.tibtech.2011.04.003.

Deep mutational scanning: assessing protein function on a massive scale

Carlos L. Araya[1] and **Douglas M. Fowler**[1]

[1] Department of Genome Sciences, 1705 NE Pacific St, University of Washington, Seattle, Washington, 98195, USA

Abstract

Analysis of protein mutants is an effective means to understand their function. Protein display is an approach that allows large numbers of mutants of a protein to be selected based on their activity, but only a handful with maximal activity have been traditionally identified for subsequent functional analysis. However, the recent application of high-throughput DNA sequencing to protein display and selection has enabled simultaneous assessment of the function of hundreds of thousands of mutants that span the activity range from high to low. Such deep mutational scanning approaches are rapid and inexpensive, with the potential for broad utility. Here, we discuss the emergence of deep mutational scanning, the challenges associated with its use, and some of its exciting applications.

Mutagenesis to understand protein function

Understanding the functional consequences of variation in protein coding sequence is of utmost importance because protein sequence dictates function, and variation in sequence often underlies pathogenesis. For example, non-synonymous mutations in protein coding sequences account for ~57% of the mutations known to be involved in human disease [1]. For many diseases, protein variation can underlie differential prognoses [2]. Genomic technologies, which are rapidly increasing in scale, are identifying large numbers of candidate mutations for a wide variety of diseases [3]. For the most common of these candidate mutations, genomic methods will likely succeed in uncovering functional significance. However, for the large majority of mutations, direct measurement will likely be required to determine their functional consequence. Furthermore, the field of protein engineering is increasingly important both for drug development and for many biotechnological applications. Effective protein engineering requires an understanding of how sequence impacts function. Thus, from both a medical and an engineering perspective, there is an acute need for better insight into how a protein's sequence relates to its function.

Mutagenesis of protein-coding sequences is a powerful way to characterize protein sequence-function relationships. Coupled to functional assays, mutagenesis has enabled the targeted and systematic probing of amino acid substitutions for a wide range of protein activities [4–6]. For example, the alanine scan involves the individual replacement of a set of residues with alanine to assess the importance of each of these residues for function [7].

Corresponding author: Fowler, D.M. (dfowler@u.washington.edu).

Attempts to more fully characterize the sequence-function relationships of a protein by evaluating large numbers of mutations have been hampered by the relatively low throughput of traditional biochemical methods, which necessitate production and purification of individual mutants. Higher-throughput methods such as protein arrays are useful, but at present are limited to hundreds or a few thousand proteins [8–10]. Protein display technologies facilitate the creation and selection for function of large libraries of protein mutants (Box 1), and have traditionally been used to select a few highly functional variants from heavily mutagenized libraries. Protein display has also been adapted to systematically interrogate protein function [11]. For example, display of a library of human growth hormone alanine mutants facilitated a shotgun alanine scan in which all alanine mutants were evaluated in parallel [12]. Until recently, such approaches have been limited to hundreds of mutants by the requirement for DNA sequencing [13].

Box 1

Protein display

Protein display refers to a diverse collection of methods that present very large numbers (10^6–10^{12}) of unique protein sequences in a format that preserves a physical link between each protein and its encoding nucleic acid sequence [13]. An initial library of proteins, or variants of a single protein, is used in a selection for a function of interest (e.g. for binding to a target), often with repeated rounds. After selection is completed, DNA sequencing of the selected library clones reveals the optimized sequences. Thus, protein display has traditionally been used to select a small number of highly optimized sequences from a diverse library. Many different protein display methods have been developed, each with its own strengths and weaknesses.

In phage display, the proteins are displayed as in-frame fusions to a phage coat protein [62]. The DNA encoding the displayed protein resides in the viral genome, packaged within the phage. Phage display systems can be used to create and evaluate libraries of up to 10^{12} clones. The commonly used M13 phage display system requires displayed proteins to be secreted from the bacteria used to produce the phage, restricting its use to short peptides [13]. The T7 phage system enables intracellular phage assembly, facilitating the display of large, multimeric proteins [63].

Bacterial [64] and yeast display [65] are conceptually similar to phage display, with the protein of interest displayed on the surface of the cell and the encoding DNA inside. These methods are more limited in the number of clones that can be displayed (~10^{10} for bacteria, ~10^6 for yeast). However, the displayed protein folds intracellularly and, in the case of yeast display, can be post-translationally modified. Additionally, these methods enable the use of fluorescence activated cell sorting (FACS) for selection. FACS facilitates quantitative screening and reduces background [13]. A cell-free alternative, ribosome display, employs *in vitro* transcription and translation of a library of mRNAs to generate polysomes displaying proteins of interest [66]. Selection is followed by reverse transcription of the attached mRNA molecules. Ribosome display has the central advantage of not requiring a cloning step, thereby generating very large libraries of displayed proteins (>10^{12}).

Over the past several years, a series of DNA sequencing technologies, collectively termed high-throughput sequencing (HTS), have been developed. These technologies acquire sequence information from millions of DNA templates in parallel (Box 2). This enormous leap in scale has made practical the use of DNA sequences as quantitative read-outs of molecular phenomena. For example, gene expression, protein-nucleic acid interactions, chromatin accessibility, chromosome conformations, translation rates, RNA modifications,

and other assays have all been adapted to make use of counting DNA sequence reads as output [3, 14–23].

Box 2

High-throughput sequencing technologies

High-throughput DNA sequencing (HTS) delivers gigabases of sequencing data at a low cost per base [39]. To achieve this feat, HTS generates millions of short sequences (50–500 bp) from solid-phase arrays of DNA molecules simultaneously. Currently available platforms include Illumina, Roche/454, ABI/SOLiD, Polonator, Helicos, and Pacific Biosciences. Sequencing proceeds by synthesis, where base incorporations are monitored by fluorescence, or in the case of pyrosequencing, by luminance [39]. The sample preparation, surface chemistry, template arraying, sequencing biochemistry, signal chemistry, ploidy (single molecule versus multiple molecule), imaging techniques, scale of miniaturization, and signal processing algorithms differ among platforms. These differences result in a diverse range of read-lengths, error rates, and read numbers produced per instrument run amongst these platforms [39].

Sequencing by synthesis strategies can be grouped into cyclic reversible termination (CRT), single-nucleotide addition (SNA), sequencing by ligation (SBL), and real-time sequencing approaches [39]. The most common of these approaches, CRT, underlies the Illumina platform and employs fluorescent reversible terminators to monitor nucleotide incorporation, producing large numbers of short reads (~100 bp). In CRT, synthesis proceeds in cycles of single-base extension with a fluorescent nucleotide terminator, imaging, cleavage of the terminator and fluorescent moiety, and washes. In sequencing by ligation (SBL), DNA ligases mediate extension of the nascent chain with fluorescent degenerate oligonucleotide sequences of fixed length flanking query positions [67]. The SBL methodology underlies the ABI SOLiD and Polonator platforms, which produce large numbers of short reads (26–50 bp).

Single nucleotide addition (SNA) approaches avoid modified bases; instead they directly monitor nucleotide incorporation. For example, in pyrosequencing, bioluminescence is used to measure the release of inorganic pyrophosphate (PPi), which is released upon nucleotide addition [68]. This methodology, commercialized by Roche/454, allows longer reads (~400 bp) than most other systems but produces lower numbers of reads per run [69].

Recently, significant progress has been demonstrated in the real-time analysis of DNA sequences polymerized by single polymerases [47]. In the Pacific Biosciences platform, this approach enables rapid acquisition and expanded read-lengths (~900 bp), but suffers from comparatively high error rates [47].

The data sets produced by HTS can be up to 50 Gb per instrument run, imposing significant computational challenges. These challenges range from the infrastructure requirements to maintain and process such volumes of data to the analysis of the information content itself. For an excellent review of current sequencing technologies, see Metzker *et al.* [39].

Recently, HTS has been coupled to protein display, making it possible to track the fitness of massive (~10^5) collections of mutants of a protein (variants) during selection for function (Figure 1). HTS quantifies the abundance of each variant in the input diversity library as well as in libraries prepared at various points during the selection process. The selection strength applied can be carefully chosen to avoid collapsing the diversity of the library excessively. Variants bearing beneficial mutations are selected for and consequently become

enriched after selection, whereas variants bearing deleterious mutations become depleted. We call this approach “deep mutational scanning.” Deep mutational scanning is of general utility, and can be applied to examine sequence-function relationships in proteins [24], RNAs [25], promoters [26], and other DNA-encoded elements [27]. In this review, we discuss the development of deep mutational scanning to analyze very large numbers of protein variants, and we subsequently consider some of the challenges and opportunities that lie ahead.

High throughput sequencing and protein display

Early studies established that HTS could be used to measure the frequency of sequences in protein display libraries [28–31]. For example, Di Niro and colleagues used phage display in combination with HTS to examine the binding to transglutaminase 2 of phage-encoded proteins derived from mRNA from different tissues [29]. Proteins of interest were identified by their abundance in the library after selection; 9 of 10 highly abundant clones were validated as strong transglutaminase 2 binders using an ELISA assay. Dias-Neto and colleagues focused on establishing the viability of using HTS to exhaustively sample a phage library consisting of protein fragments derived from surgical biopsies [28]. They sequenced 3840 individual phage clones using Sanger sequencing and compared these to the 319,361 sequences they obtained by HTS using the Roche/454 platform. They found that HTS did not alter GC content, codon usage or amino acid frequencies reported for the encoded protein fragments.

Several other examples involved one of the most prominent uses of phage display: identifying high-affinity antibodies. Ge and coworkers developed an amplification-free method for constructing and refining synthetic immunoglobulin libraries using DNA oligonucleotide assembly followed by selection for in frame inserts [31]. To demonstrate the utility of their approach they characterized >250,000 clones constructed by their method using HTS, which illustrates the applicability of HTS for examining large libraries of protein variants. Ravn and colleagues used HTS to facilitate the identification of high-affinity antibody sequences in a phage display library of $>10^7$ antibody fragments (scFv), mutagenized in an important loop region (CDR3) [30]. Sequencing of phage libraries after selection for binding to an epitope enabled an enumeration of high-frequency sequences. They tested six high-frequency sequences and found that all had enhanced affinity for the epitope. This work showed that high affinity antibody fragments could be identified with less up-front screening, illustrating how HTS approaches can vastly improve the identification and characterization of novel antibodies. HTS has also been employed to examine the repertoire of antibodies present within living organisms [32–35]; interested readers are directed to Fischer’s useful commentary [36].

These studies illustrated that protein display experiments, which traditionally start with a highly diverse library and converge on a small number of proteins bearing the function of interest, can be greatly enhanced with HTS. By demonstrating that HTS can be used to quantify the abundance of sequences in protein display libraries, these experiments prefigured deep mutational scanning. They also outlined many of the inherent challenges in combining HTS and protein display, particularly the biases in sequence representation and the relatively high error rate associated with HTS.

Measuring sequence-function relationships

Recently, HTS has been employed to examine the effect of mutations on the function of peptides and proteins. Ernst and colleagues [37] used deep mutational scanning to explore the peptide binding preferences of 22 distinct members of the PDZ domain family. PDZ domains are peptide recognition modules approximately 90 amino acids in length. Each of

the 22 PDZ domains was evolved *in vitro* to have altered peptide-binding specificities. These evolved domains were used to select peptide ligands from a phage displayed ~10^{11} member library of random heptapeptides. By employing a barcoding strategy, they sequenced all 22 selections simultaneously.

The resulting 44,097 sequences represented 25,566 unique peptides divided amongst the 22 PDZ domains. Peptide binding sequences for each of the 22 domains were analyzed using a machine-learning algorithm to uncover each domain's specificity profile. Each specificity profile represents the peptide binding preferences of its PDZ domain. Interestingly, many of the domains exhibited multiple specificity profiles, showing that some domains could accommodate several distinct peptide sequences. Comparison of the evolved domains specificity profiles to profiles of natural PDZ domains revealed that the evolved PDZ domains were less specific than their natural counterparts. As expected, the evolved PDZ domains also bound their target peptides with higher affinity. The authors infer that the reduced selectivity of their evolved PDZ domains arises from the fact that they have fewer interactions with peptide side chains than do natural domains.

The authors examined their data for evidence of correlated changes between evolved PDZ domains and their cognate peptides. They found evidence of significant correlation at four positions within the PDZ domain. In each case, mutations altering the charge of the PDZ domain introduced favorable electrostatic interactions with its peptide ligand. Thus, the use of HTS to simultaneously profile the binding preferences of 22 distinct evolved PDZ domains enabled the authors to gain insight into the co-evolution of a protein-peptide interaction.

In a set of experiments explicitly designed to investigate protein sequence-function relationships, Fowler and colleagues studied the interaction between the hYAP65 WW domain and its cognate polyproline-rich peptide [24]. In the PDZ experiments discussed above, the functional capacity of a given variant (i.e. variant fitness) was defined based only on the variant's abundance after selection. Ignoring the abundance of each variant in the input library hinders accurate assessment of variant fitness because biases in library construction mean that variants are generally not equally abundant in the input library. For example, abundances in the input library of WW variants ranged over four orders of magnitude. To ameliorate this problem, the authors used the abundance of each variant before and after selection to calculate an enrichment ratio. The use of enrichment ratios corrected for the nonuniform abundance of variants in the input library and resulted in more accurate assessment of the fitness of each variant.

To facilitate the assessment of variants bearing deleterious mutations, an input library of low complexity with an average of 2 mutations per variant was used [24]. Using a low-complexity input library (i.e. one with fewer variants) ensured a large number of copies of each variant. Thus, the decrease in abundance of variants bearing deleterious mutations could be characterized with statistical confidence. Examination of the enrichment and depletion of ~10^5 variants of the WW domain through selection allowed the construction of a nearly complete map that describes the effects of mutations within the WW domain. The effect of 53 out of 56 mutations previously measured by standard biochemical methods were corroborated by the data generated by deep mutational scanning.

At the primary structure level, mutagenesis can reveal the precise location of catalytic residues as in classical alanine scans [7]. In the context of secondary and tertiary structural information, mutational preferences can highlight regions that are generally intolerant of or favorable to mutation. The WW and PDZ domain studies discussed above illustrate how deep mutational scanning, used in concert with domain and structural information, reveals

critical residues and regions in proteins and in turn aids in understanding the molecular determinants of fitness. Thus, by massively parallelizing a fitness assay, deep mutational scanning enables rapid large-scale measurements of protein fitness and offers an unprecedented view of protein sequence-function relationships.

Enhancing quantitation

Massive parallelization of measurement often results in low accuracy, and the relatively low quality of HTS reads presents a potentially significant problem in faithfully identifying variants within deep mutational scans. For example, the Illumina platform has an average per-base error rate of ~1%, so a 75-base read would be expected to contain one incorrectly identified point mutation [38]. These errors could introduce a large amount of noise when estimating variant fitness, particularly for variants present at low copy number. The error rates for other platforms vary, but all are high enough to be problematic [39]. Solutions to high error rates have been implemented, including using a paired-end approach to enhance fidelity or exclusion of sequences observed fewer than a certain number of times [24, 28, 29]. An effective solution to sequencing error requires that each library member bear a unique barcode sequence [26, 40]. Barcodes can be used to identify replicate instances of each library member. Examination of replicates enables error correction [40, 41]. Furthermore, barcoding each variant would enable explicit quantitation of and correction for library biases introduced by amplification steps such as PCR.

Another major challenge in using deep mutational scanning is to enhance the accuracy of the derived fitness measurements. Work using HTS to study libraries of DNA variants that are selected for binding to a protein offer useful methods for accomplishing this goal. Studies by Jolma and colleagues [42] and Zhao and colleagues [43] demonstrate the utility of explicitly tracking DNA variants through sequential rounds of selection, which can facilitate correction for artifacts such as non-specific carryover and saturation to greatly enhance the accuracy of the fitness estimate for each variant. For example, the non-specific carryover rate can be estimated using variants containing early stop codons, as these should be completely non-functional. The estimated non-specific carryover rate can then be used to adjust variant fitness estimates [42]. The application of these models to sequential selections of protein libraries could significantly improve the accuracy of fitness estimates. In addition, incorporation of error estimates that take into account the number, diversity, and quality of sequences from which mutation effects are derived should greatly improve reliability.

The generation of variant libraries is another area that is ripe for improvement. Random mutagenesis, either enzymatically or by chemical synthesis, produces variants containing different numbers of mutations. The result is a library that contains some variants with fewer than the desired number of mutations and some with more than the desired number. Furthermore, the constraints of the genetic code dictate that, given a particular wild type sequence, some amino acid changes are more likely than others. These biases reduce the complexity of the library that can be employed. Thus, cost-effective library generation methods that allow precise sequence tailoring and unbiased representation of variants are needed. One possible solution is trinucleotide-based oligonucleotide synthesis, which employs pre-synthesized, codon-sized nucleotide trimers that eliminate bias owing to the constraints of the genetic code [44]. Another promising approach relies on the application of custom DNA arrays and synthetic gene-assembly [31, 45, 46], which could enable exact specification for relatively large libraries, eliminating the problems associated with random mutagenesis.

The application of deep mutational scanning to proteins of increased size will necessitate longer sequencing reads to cover the mutated region as well as higher read numbers to

account for larger sequence space as length increases. Advances in HTS technology could resolve this problem by increasing read lengths dramatically [47]. Currently available alternative approaches to increase effective read length, such as subassembly, could also be of great value [40].

Understanding high-dimensionality mutagenesis data

Previous mutagenesis studies generally yielded fitness data on, at most, hundreds of variants. Deep mutational scanning has been used to analyze 10^4–10^5 variants [24, 37]. Of these, only a small number were single mutants, and the rest were multiple mutants. Fitness data of multiply mutated variants are of high dimensionality, so their visualization and analysis present a challenge. Several approaches have been employed, including motifs, maps and landscapes (Figure 2). Sequence motifs illustrate the average representation of amino acids (or, more commonly, nucleotides) on a per-position basis within a set of sequences (Figure 2a). In a selected library, motifs constitute a straightforward view of beneficial amino acids at each position. Single motifs discard much of the detail generated by deep mutational scanning, because all variants are averaged and deleterious amino acids are not visualized. However, computational approaches that produce multiple motif profiles offer a promising way to overcome this limitation [37]. Maps entail a set of sequence-function relationships representing the fitness effect of every possible single mutation (Figure 2b). As such, maps provide a detailed view of mutational preferences that includes deleterious, neutral and beneficial mutations. Maps can be generated from single-mutant data alone, or by averaging data from mutations occurring on multiply mutated backgrounds [24]. However, maps ignore dependencies among interacting residues in multiply mutated variants, which means that unique combinations of mutations that confer strong fitness phenotypes are not visualized.

Landscapes constitute a third approach to understanding sequence-function relationships generated by deep mutational scanning (Figure 2c). Landscapes display fitness as a function of sequence distance, where peaks in the sequence-function landscape at increasing distances from a wild type reveal functionally analogous but increasingly unrelated variants. Deep mutational scanning using currently available sequencing technology can derive fitness measurement for ~10^5 individual protein variants. A 50 amino acid sequence has 10^3 single mutants, 10^5 double mutants and 10^8 triple mutants. Thus, a deep mutational scan of a small protein domain could be used to construct a nearly complete double mutant landscape, or a much more sparsely populated higher-order landscape.

Constructing a landscape requires that the relatedness of each sequence be calculated. One of the most common measures of distance is the number of mutations a variant has relative to a master sequence (the Hamming distance) [48]. More advanced methods of calculating distance are available [49, 50], but are beyond the scope of this review. Transformations of high-dimensional sequence data to a lower-dimensional array can be problematic [29, 51–53], but landscapes are attractive visual representations because the ability to identify separate peaks allows motifs to be generated for distinct regions of sequence space. Functionally analogous peaks, distinct in sequence space, could uncover functional redundancy in genomes that would be inaccessible to conventional homology-based searches. Several HTS-based studies illustrate the utility of landscape representations. These include a study of ~10^{13} variants of the class II ligase ribozyme [25] and a study of the DNA binding preferences of several transcription factors [50]. Motifs, maps and landscapes differ in their utility, computational requirements and illustrative power. Ultimately, new computational approaches are needed, which could include interactive tools or network-style projections of interacting positions.

Outlook and conclusions

The advances described here enable an exciting array of experiments. By obtaining fitness measurements for hundreds of thousands of proteins simultaneously, deep mutational scanning will facilitate experimental investigation of several long-discussed features of proteins including epistasis and fitness landscape topology. Smaller-scale studies have shown that most mutations have deleterious effects on protein fitness, and that as proteins accumulate mutations their fitness decays [29, 52, 54, 55]. Deep mutational scanning experiments provide experimental validation that the vast majority of mutations are deleterious [24].

Deep mutational scanning could contribute substantially to the study of how multiple mutations interact (epistasis). Epistatic interactions can either be negative, when the effect of combined mutations is more deleterious than an independent model would predict, or positive, when the effect of combined mutations is more beneficial than the model would predict. Epistasis has important consequences for our understanding of protein evolution, as it implies that lineages leading to fitter variants can require trajectories through non-optimal intermediates. Therefore, the rate of protein evolution is likely to vary along different evolutionary trajectories, with this variance introducing error into estimates of evolutionary distances in protein families [56]. Furthermore, mutations that stabilize the folded structure of a protein can have powerful positive epistatic interactions with other adaptive mutations [53].

Although protein epistasis has been studied extensively, experimental limitations have restricted the number of epistatic interactions tested. In deep mutational scanning experiments with a WW domain, the fitness of 10,192 variants containing two mutations in the domain could be predicted from single mutant data using a simple product model with ~70% accuracy [24]. These results illustrate the power of deep mutational scanning to provide a large number of experimental epistasis measurements. With refined quantitation, deep mutational scanning could be used to measure epistatic interactions between residues at previously inaccessible levels of detail, thereby enabling more accurate epistasis models to be constructed.

Computational analysis of protein folding mechanisms, protein structure and protein function represents a frontier in protein science. Despite their recent astonishing advances, computational methods can be improved significantly. Computational methods benefit from training and validation data sets, but these have been limited in scope. Deep mutational scanning offers the opportunity to generate empirical fitness measurements for thousands to millions of protein variants, providing validation and training sets orders of magnitude larger than those currently available. Thus, deep mutational scanning could engender a rich and productive symbiosis between experiment and computation, thereby accelerating innovations in drug and protein design.

The ability to catalog genomic alterations associated with disease has greatly accelerated with the advent of HTS, but technologies for measuring the effect of these alterations have been much more limited. For example, the Cancer Genome Atlas and other studies have identified many genes that are mutated in cancer [57–60]. However, the abundance of passenger mutations unrelated to cancer confounds these efforts and hampers our capacity to gauge the effect of mutations within these candidate genes. Mutations also frequently underlie the failure of targeted cancer therapeutics [61]. Deep mutational scanning of cancer-related proteins can enable an understanding of how mutations affect protein function and inhibitor efficacy, enhancing the meaning and clinical utility of genomic efforts. More

generally, data derived from deep mutational scanning of disease-linked proteins could be used to guide prognosis and prevention for individuals with deleterious variants.

Deep mutational scanning offers the capacity to parallelize biochemistry, measuring fundamental biochemical properties such as substrate affinity, catalytic activity, and folding energy on a massive scale. Furthermore, we anticipate that deep mutational scanning will be extensively applied *in vivo*. However, deep mutational scanning *in vivo* may require improvements to address the effects of cell-by-cell variances in protein expression and stochastic fluctuations in library composition associated with transformation bottlenecks. Understanding more complex phenotypes, such as fitness effects arising from variation in gene regulatory activities, brings up numerous challenges in the interpretation of genotype-phenotype relationships. Nevertheless, we envision that the use of sequencing as a way of quantifying changes in DNA-encoded populations will continue to grow.

Acknowledgments

We would like to thank Stan Fields, Jay Hesselberth and Alan Rubin for helpful discussion and comments regarding the manuscript. This work was supported by the National Institutes of Health (F32GM084699 to D.M.F.)

Glossary

Protein display	A collection of techniques that preserve the link between a protein and its encoding nucleotide sequence, enabling the creation of libraries consisting of a large number of different proteins (see Box 1)
High-throughput sequencing (HTS)	A set of sequencing technologies that parallelize the acquisition of DNA sequencing data, generating gigabases of sequence information at low cost (see Box 2)
Barcoding	Barcoding entails tagging each DNA element in a library with a DNA barcode. Tagging facilitates combining DNA libraries for sequencing with later deconvolution based on the barcode identity
Paired-end sequencing	The acquisition of DNA sequence reads from both strands of the same DNA molecule. Paired-end sequencing can be overlapping, resulting in twofold coverage of the same DNA sequence from a single paired-end read
Motif	A representation that expresses, at each position within the motif, the relative abundance of each amino acid or nucleotide
Sequence-function map	A representation of the fitness effects of making all possible single amino acid substitutions within a protein. Sequence-function maps can be generated from single mutant variants or can be created by averaging the effects of mutations occurring in multiply mutated variants
Sequence-function landscape	A representation of fitness as a function of sequence, where increasing distance in the landscape corresponds to increasing distance in sequence space
Variant	A unique protein or DNA sequence generated by mutagenesis from a wild type sequence. A variant can be present in one or many copies in a library

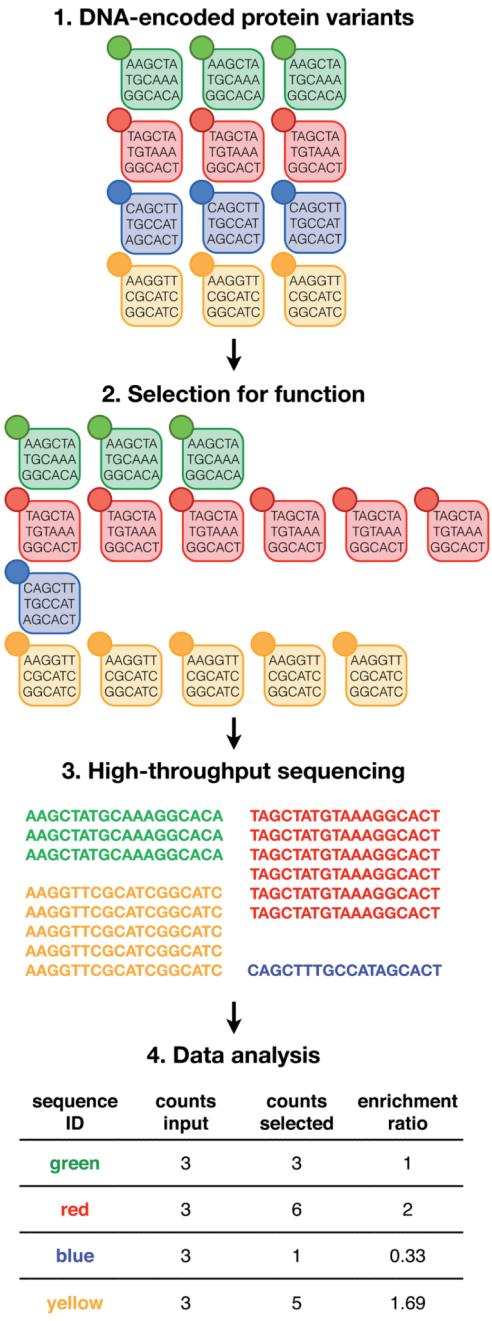
Deep mutational scanning	A technique that assesses the fitness of a large number of mutants of a protein (or other DNA-encoded functional elements) simultaneously using high-throughput DNA sequencing in combination with selection for function

References

1. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
2. Olivier M, et al. The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clin Cancer Res. 2006; 12:1157–1167. [PubMed: 16489069]
3. Ng S, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]
4. Pal G, et al. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. J Biol Chem. 2006; 281:22378–22385. [PubMed: 16762925]
5. Matouschek A, et al. Mapping the transition state and pathway of protein folding by protein engineering. Nature. 1989; 340:122–126. [PubMed: 2739734]
6. Sidhu SS, Koide S. Phage display for engineering and analyzing protein interaction interfaces. Curr Opin Struct Biol. 2007; 17:481–487. [PubMed: 17870470]
7. Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science. 1989; 244:1081–1085. [PubMed: 2471267]
8. MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. Science. 2000; 289:1760–1763. [PubMed: 10976071]
9. Zhu H, et al. Global analysis of protein activities using proteome chips. Science. 2001; 293:2101–2105. [PubMed: 11474067]
10. Ramachandran N, et al. Next-generation high-density self-assembling functional protein arrays. Nat Methods. 2008; 5:535–538. [PubMed: 18469824]
11. Sidhu SS, Kossiakoff AA. Exploring and designing protein function with restricted diversity. Curr Opin Chem Biol. 2007; 11:347–354. [PubMed: 17500026]
12. Weiss GA, et al. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc Natl Acad Sci USA. 2000; 97:8950–8954. [PubMed: 10908667]
13. Levin AM, Weiss GA. Optimizing the affinity and specificity of proteins with molecular display. Mol Biosyst. 2006; 2:49–57. [PubMed: 16880922]
14. Ng SB, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010; 42:790–793. [PubMed: 20711175]
15. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448:553–560. [PubMed: 17603471]
16. Licatalosi DD, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature. 2008; 456:464–469. [PubMed: 18978773]
17. Nagalakshmi U, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320:1344–1349. [PubMed: 18451266]
18. Core LJ, et al. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. [PubMed: 19056941]
19. Sultan M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321:956–960. [PubMed: 18599741]
20. Valouev A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008; 5:829–834. [PubMed: 19160518]
21. Ingolia NT, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–223. [PubMed: 19213877]
22. Duan Z, et al. A three-dimensional model of the yeast genome. Nature. 2010; 465:363–367. [PubMed: 20436457]

23. Zaranek AW, et al. A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. PLoS Genet. 2010; 6:e1000954. [PubMed: 20531933]
24. Fowler DM, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010; 7:741–746. [PubMed: 20711194]
25. Pitt JN, Ferre-D'Amare AR. Rapid construction of empirical RNA fitness landscapes. Science. 2010; 330:376–379. [PubMed: 20947767]
26. Patwardhan RP, et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol. 2009; 27:1173–1175. [PubMed: 19915551]
27. Cho M, et al. Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. Proc Natl Acad Sci USA. 2010; 107:15373–15378. [PubMed: 20705898]
28. Dias-Neto E, et al. Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. PLoS One. 2009; 4:e8338. [PubMed: 20020040]
29. Di Niro R, et al. Rapid interactome profiling by massive sequencing. Nucleic Acids Res. 2010; 38:e110. [PubMed: 20144949]
30. Ravn U, et al. By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res. 2010; 38:e193. [PubMed: 20846958]
31. Ge X, et al. Rapid construction and characterization of synthetic antibody libraries without DNA amplification. Biotechnol Bioeng. 2010; 106:347–357. [PubMed: 20198660]
32. Reddy ST, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol. 2010; 28:965–969. [PubMed: 20802495]
33. Boyd SD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009; 1:12ra23.
34. Weinstein JA, et al. High-throughput sequencing of the zebrafish antibody repertoire. Science. 2009; 324:807–810. [PubMed: 19423829]
35. Glanville J, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA. 2009; 106:20216–20221. [PubMed: 19875695]
36. Fischer N. Sequencing antibody repertoires: The next generation. MAbs. 2011; 3:17–20. [PubMed: 21099370]
37. Ernst A, et al. Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. Mol Biosyst. 2010; 6:1782–1790. [PubMed: 20714644]
38. Quail MA, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008; 5:1005–1010. [PubMed: 19034268]
39. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
40. Hiatt JB, et al. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. 2010; 7:119–122. [PubMed: 20081835]
41. Ilie L, et al. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics. 2011; 27:295–302. [PubMed: 21115437]
42. Jolma A, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 2010; 20:861–873. [PubMed: 20378718]
43. Zhao Y, et al. Inferring binding energies from selected binding sites. PLoS Comput Biol. 2009; 5:e1000590. [PubMed: 19997485]
44. Virnekas B, et al. Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res. 1994; 22:5600–5607. [PubMed: 7838712]
45. Kosuri S, et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat Biotechnol. 2010; 28:1295–1299. [PubMed: 21113165]
46. Borovkov AY, et al. High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. Nucleic Acids Res. 2010; 38:e180. [PubMed: 20693531]

47. Eid J, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323:133–138. [PubMed: 19023044]
48. Hamming RW. Error Detecting and Error Correcting Codes. Bell System Technical Journal. 1950; 29:147–160.
49. Pitt JN, et al. SEWAL: an open-source platform for next-generation sequence analysis and visualization. Nucleic Acids Res. 2010 epub ahead of print.
50. Carlson CD, et al. Specificity landscapes of DNA binding molecules elucidate biological function. Proc Natl Acad Sci USA. 2010; 107:4544–4549. [PubMed: 20176964]
51. Carneiro M, Hartl DL. Colloquium papers: Adaptive landscapes and protein evolution. Proc Natl Acad Sci USA. 2010; 107(Suppl 1):1747–1751. [PubMed: 19805125]
52. Bershtein S, et al. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006; 444:929–932. [PubMed: 17122770]
53. Bloom JD, Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. Proc Natl Acad Sci USA. 2009; 106(Suppl 1):9995–10000. [PubMed: 19528653]
54. Azevedo RB, et al. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature. 2006; 440:87–90. [PubMed: 16511495]
55. Elena SF, Lenski RE. Test of synergistic interactions among deleterious mutations in bacteria. Nature. 1997; 390:395–398. [PubMed: 9389477]
56. Pal C, et al. An integrated view of protein evolution. Nat Rev Genet. 2006; 7:337–348. [PubMed: 16619049]
57. Ding L, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. [PubMed: 18948947]
58. Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446:153–158. [PubMed: 17344846]
59. Shah SP, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009; 461:809–813. [PubMed: 19812674]
60. Wood LD, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318:1108–1113. [PubMed: 17932254]
61. Quintas-Cardama A, et al. Mechanisms of primary and secondary resistance to imatinib in chronic myeloid leukemia. Cancer Control. 2009; 16:122–131. [PubMed: 19337198]
62. Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science. 1985; 228:1315–1317. [PubMed: 4001944]
63. Dai M, et al. Using T7 phage display to select GFP-based binders. Protein Eng Des Sel. 2008; 21:413–424. [PubMed: 18469345]
64. Francisco JA, et al. Transport and anchoring of beta-lactamase to the external surface of Escherichia coli. Proc Natl Acad Sci USA. 1992; 89:2713–2717. [PubMed: 1557377]
65. Feldhaus MJ, et al. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. Nat Biotechnol. 2003; 21:163–170. [PubMed: 12536217]
66. Mattheakis LC, et al. An in vitro polysome display system for identifying ligands from very large peptide libraries. Proc Natl Acad Sci USA. 1994; 91:9022–9026. [PubMed: 7522328]
67. Shendure J, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309:1728–1732. [PubMed: 16081699]
68. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]
69. Rothberg J, Leamon J. The development and impact of 454 sequencing. Nat Biotechnol. 2008; 26:1117–1124. [PubMed: 18846085]



sequence ID	counts input	counts selected	enrichment ratio
green	3	3	1
red	3	6	2
blue	3	1	0.33
yellow	3	5	1.69

Figure 1. Deep mutational scanning to measure protein sequence-function relationships

Deep mutational scanning takes advantage of high-throughput DNA sequencing to measure the function of variants of a protein on a massive scale. A color-coded population of DNA encoded protein variants is shown (1). Each solid circle denotes a displayed protein variant linked to its encoding DNA sequence. In this idealized input library, variants have equal representation. The library is shown after selection for function of the displayed protein (2). Variants bearing beneficial mutations increase in abundance in the selected library (e.g. red and yellow) whereas deleterious variants decrease in abundance (e.g. blue). High-throughput sequencing is performed on the selected and input libraries (3). The number of times each variant is sequenced corresponds to its abundance in the library (the example shown is for

the selected library). Sequencing data from the input and selected libraries are used to calculate an enrichment ratio for each variant (4). The enrichment ratio of a variant is a measure of its fitness.

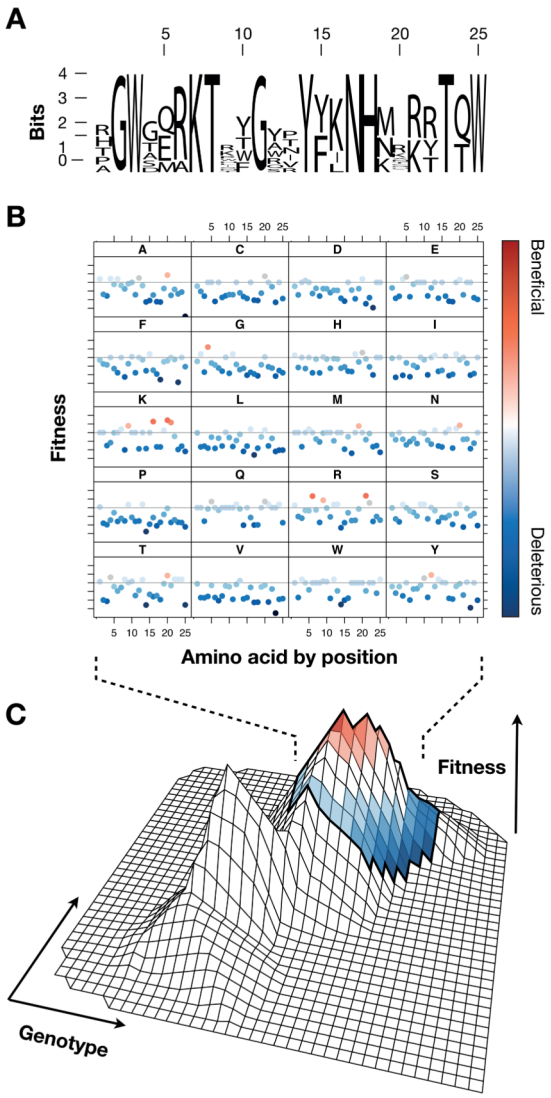


Figure 2. Motifs, maps and landscapes for visualizing sequence-function relationships

Deep mutational scanning generates a large set of sequence-function relationships; three methods for visualizing these relationships are shown. Data for a 25 amino acid deep mutational scan of the WW domain was used to create each panel (Sequence Read Archive accession SRA020603) [24]. (**a**) An amino acid motif is shown. This motif illustrates the abundance of each amino acid at each position in the selected library of variants. (**b**) A sequence-function map is shown, which was generated by calculating the fitness effect of each amino acid at every position. In the example given, fitness corresponds to the enrichment or depletion of sequences with specific substitutions during selection. Mutations are color coded from beneficial to deleterious in a red to blue color range, respectively. Gray dots indicate substitutions with neutral fitness relative to the reference sequence. (**c**) A sequence-function landscape is shown. The x- and y-axes denote measures of sequence distance (i.e. points that are close to each other represent variants with similar sequences) and the z-axis illustrates fitness. The region of the landscape within a single amino acid substitution of the reference variant corresponds to the map described in (**b**) and is colored accordingly.