Highlights

- Deep sequencing permits examination of local protein fitness landscapes
- NGS aids in engineering affinity and specificity for protein molecular recognition
- Application to enzymes is limited by paucity of functional selections
- NGS enables fundamental studies of enzyme behavior and evolution
- Gene tiling and molecular barcoding extend NGS to full-length proteins

Deep Sequencing Methods for Protein Engineering and Design

Emily E. Wrenbeck[1], Matthew S. Faber[2], Timothy A. Whitehead[1,3]taw@egr.msu.edu

[1]Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, Michigan, 48824.

[2]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, 48824.

[3]Departments of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, Michigan, 48824.

Abstract

The advent of next-generation sequencing (NGS) has revolutionized protein science, and the development of complementary methods enabling NGS-driven protein engineering have followed. In general, these experiments address the functional consequences of thousands of protein variants in a massively parallel manner using genotype-phenotype linked high-throughput functional screens followed by DNA counting via deep sequencing. We highlight the use of information rich datasets to engineer protein molecular recognition. Examples include the creation of multiple dual-affinity Fabs targeting structurally dissimilar epitopes and engineering of a broad germline-targeted anti-HIV-1 immunogen. Additionally, we highlight the generation of enzyme fitness landscapes for conducting fundamental studies of protein behavior and evolution. We conclude with discussion of technological advances.

1. Introduction

Researchers have been engineering proteins for almost 4 decades. Early endeavors involved generation of a handful of point mutations followed by low-throughput assays for function; the ‘search space’ a protein scientist could feasibly explore was miniscule.

As demonstrated by the seminal works of Fowler et al. [1] and Hietpas et al. [2], the advent of next-generation sequencing (NGS) has presented protein engineers with the ability to economically observe *entire populations* of molecules before, during, and after a high-throughput screen or selection for function (HTS) (**Figure 1**). A typical NGS run provides sufficient sequencing data to permit the study of tens of thousands of protein variants. Thus, when coupled to HTS, NGS significantly expands the accessible mutational search space. In this way a researcher can test all possible point mutations or combinations of mutations, for example, and remove the duty of having to design small focused libraries that may miss unpredictable beneficial mutations. As a testimonial to the accessibility of these methodologies, experiments can be performed in a beginning graduate-level course [3].

The intent of this review is to highlight examples where deep sequencing has been applied in different areas of protein engineering and design. As such, we will not provide a comprehensive review of directed evolution or of deep mutational scanning (excellent reviews can be found here [4,5]). We will discuss the use of NGS for engineering protein molecular recognition, membrane proteins, and enzymes, highlight recent technological advances, and offer a perspective on the shape of the field over the next several years.

1. Engineering Protein Molecular Recognition

Dozens of studies over the past five years have used deep sequencing to identify and engineer protein-ligand interactions. Rapid adoption of deep sequencing by this field is a direct result of mature display-based technologies that can be used to screen very large initial libraries. For example, in the study of protein-protein binding interactions a library of protein variants can be displayed on the surface of yeast using yeast surface display (**Figure 1**). Yeast cells are labeled with a fluorescently conjugated binding partner, and FACS can be used to screen cells by fluorescence intensity.

1.1. Deep sequencing for screening protein binder libraries

NGS is now frequently used in the evaluation of synthetic or natural libraries to identify antigen-specific binders. Advances in pairing V_H and V_L sequences from individual B cells [6] allows one to identify antigen-specific antibodies directly from sequencing, including panels of antibodies targeting Ebola virus [7] and ricin [8]. Methodological details and limitations associated with identification of rare clones and evaluation of library diversity are presented in a recent review [9].

As an emerging area, engineers now use NGS to refine protein binder libraries . In a notable advance, Woldring et al. screened a hydrophilic fibronectin domain library to bind various protein targets [11]. The researchers exploited the site-specific amino acid preferences from an initial library to develop a more focused second library depleted in mutations at the periphery of the binder paratope. Compared to other libraries, this library design afforded far superior performance in isolation of high affinity, stable binders.

1.1. Paratope optimization for affinity and specificity

NGS can be used to rapidly improve the affinity and specificity of the binding paratope (**Figure 2**) [12,13]. A crucial advantage enabled by NGS is the ability to discriminate very small beneficial changes in binding - on the order of 0.1 kcal/mol or about a 20% improvement in dissociation constant. These small-scale beneficial mutations can be additive, allowing one to “leapfrog” over potential affinity maturation bottlenecks by combining mutations.

Whitehead et al. provide the first example of paratope engineering for affinity and specificity using deep sequencing [14]. The researchers screened a comprehensive single-site saturation mutagenesis library of two *de novo* designed Influenza Hemagglutinin (HA) binders against H1 and H5 HA subtypes. Engineering specificity was demonstrated by comparing site-specific preferences for H1 to the H5 subtype. A single point mutation was identified that gave over a 30-fold specificity switch from the parental designed protein. For affinity maturation, site-specific preferences were encoded into a second library and sorted to improve affinity against both subtypes by approximately 25-fold. The affinity of one designed HA binder, HB36.6, was further improved against seven diverse HA subtypes. HB36.6 showed prophylactic and therapeutic efficacy against lethal challenge of pandemic Influenza in a BALB/c mouse model [15].

Deep mutational scanning approaches have been extended to affinity mature antibodies [16,17]. In an impressive demonstration, Genentech scientists engineered a dual action Fab for high affinity for two unrelated proteins simultaneously [18]. The group used phage display to profile a single and triple site saturation mutagenesis library of a Fab with low nanomolar binding to Ang2 and VEGF. NGS revealed significant site-specific amino acid preferences for each of the two binding paratopes. The researchers combined mutations shown to improve affinity on at least one target and not negatively impact binding on the other target, thus engineering five different sub-nanomolar dual-affinity Fabs.

The apotheosis of deep mutational scanning to identify high affinity binders with defined specificity comes from Jardine et al. [19], who engineered an HIV immunogen that can be recognized by B cell precursors to broadly neutralizing anti-HIV antibodies. Starting with a designed outer domain of the gp120 protein from HIV, they screened a 58-residue site saturation mutagenesis library against 18 germline-reverted and 11 VRC01-class broadly neutralizing antibodies. Information obtained from the scan was used to encode a second library that was screened against the same antibody panel. One variant showed dramatically improved binding to all antibodies in the panel and could bind naïve B cells in full human repertoires.

Binding surface optimization is not limited to protein-protein binders, provided that there is a suitable HTS. Tinberg et al. used yeast display coupled to NGS to affinity mature a computationally designed anti-steroid binder [20]. Raman and colleagues used an *in vivo* fluorescent reporter coupled to FACS (**Figure 1**) to engineer the *E. coli* allosteric transcription factor LacI to recognize four different non-metabolizable inducers, including sucralose [21].

1.1. Epitope mapping

An important consideration for the antibody engineer is the identification of the binding epitope. Three recent publications used yeast surface display, site-saturation mutagenesis, FACS, and deep sequencing to identify conformational epitopes for diverse antigenic targets on the order of weeks [22–24]. Doolan and Colby determined epitope regions on prions recognized by conformational-specific antibodies [22]. Van Blarcom et al.

performed epitope mapping for a panel of antibodies against the alpha toxin from methicillin-resistant *Staphylococcus aureus* [23]. Kowalsky et al. automated and improved the speed of epitope identification for three different antigens [24].

1. Membrane Protein Engineering

There are few examples of deep sequencing-enabled membrane protein engineering. In the best example, Plückthun and colleagues screened a near-comprehensive single point mutant library of G protein-coupled receptor (GPCR) rat neurotensin receptor 1 for enhanced heterologous expression, a proxy for protein stability. The library was expressed in the periplasm of *E. coli* and sorted by FACS using a fluorescently conjugated agonist as a probe [25]. NGS was used to quantify variants in the input library and the enriched FACS selected libraries, and hits identified in the initial library were combined, resulting in variants that express at up to 50-fold higher levels in *E. coli* compared with the wild-type GPCR. Each stability-enhancing mutation contributed a small amount of the overall stability to the protein [26]. Notably, the structure of an engineered GPCR was solved [27], suggesting a general directed evolution strategy of stabilizing membrane proteins for X-ray crystallography structure determination.

1. Enzyme Engineering

In contrast to protein-ligand interactions, the complex and diverse nature of enzyme function has made it challenging to develop robust, sensitive, and generalizable functional screens. As such, far fewer examples of deep sequencing-assisted enzyme engineering exist in the literature (**Table 1**).

1.1. High-throughput screening and selection for enzyme function

The primary strategy for functional selection of enzymes is to tether enzymatic function to the growth and/or survival (fitness) of a host organism. One type of competitive growth selection is to provide a substrate that the enzyme must catabolize as the sole source of an essential element for growth (carbon, nitrogen) (**Figure 1**). Thus, variants enabling higher flux through and enzyme permit faster growth rates and become enriched in the population. Klesmith et al. performed deep mutational scanning of levoglucosan kinase, where levoglucosan was fed as the carbon source [28]. Similarly, Wrenbeck et al. performed deep mutational scanning on amiE, an aliphatic amidase from *Pseudomonas aeruginosa*, by feeding amides as the nitrogen source [E. E. Wrenbeck et al., unpublished]. Antibiotic resistance genes also provide straightforward targets for competitive growth selections. Indeed, these represent 4/9 published enzyme scans (**Table 1**) [29–31]. In summary, high-throughput screens or selections that are *generalizable* are desired, yet the incredible diversity of enzyme function makes their development a critical challenge for the field.

1.1. From fitness landscapes to enzyme engineering

Deep mutational scanning experiments afford a richness of knowledge of ‘hits’. However, efficiently utilizing ambiguous ‘fitness values’ to inform enzyme design is still a significant challenge. To avert this challenge, van der Meer et al. performed over 4000 assays to generate ‘mutability landscapes’ of a tautomerase enzyme for its expression, Michael-type activities on multiple substrates, and characterization of its enantioselectivity, and used this information to design a novel enantioselective Michaelase [32].

How does one intelligently combine hits to achieve a given design goal? One approach is to biophysically characterize beneficial mutations. For example, Klesmith et al. performed deep mutational scanning of levoglucosan kinase to identify mutations that improved fitness through improved flux of levoglucosan conversion. They characterized a set of beneficial mutations for activity and thermodynamic stability and used this information to generate designs, one of which had greater than 24-fold improvement in activity and 7°C increase in apparent melting temperature [28]. An alternative approach is to generate multiple fitness landscapes under different conditions (concentration and identity of substrate, temperature, etc.) and use differential analysis to generate designs. To that end Melnikov et al. performed deep mutational scanning of APH(3')II, an enzyme responsible for aminoglycoside antibiotic resistance, with several antibiotics at different concentrations and generated designs with orthogonal activities [33].

Datasets from deep mutational scanning can be used to probe the fundamental nature of enzyme behavior and can be used to ask questions related to evolutionary trajectories, rigorously testing theories gleaned from over two decades of directed evolution experiments. Steinberg and Ostermeier analyzed fitness effects for TEM-15 β-lactamase under varying environmental conditions and found that negative selections were able to bridge access to the highest fitness peaks [34]. Wrenbeck et al. performed deep mutational scanning of an aliphatic amidase on three substrates and found that specificity-determining mutations were distributed throughout the protein sequence and structure rather than located near the active site [E. E. Wrenbeck et al., unpublished].

1. Methodological Advances and Current Limitations

1.1. Mutagenic Library Preparation

Consider a protein of a typical length of 300 residues. A library comprising every possible single or double point mutation would contain 6x10^3 or 3.6x10^7 sequences, respectively. Similarly, a library with simultaneous saturation mutagenesis at four defined positions contains 1.6x10^5 sequences. For a typical experimental workflow there are 10^6-10^7 quality-filtered DNA reads, and accurate estimation of variant frequencies occurs above a statistical background of ~100 sequence reads per variant [35,36]. Dividing the number of sequences from a NGS run by the minimum number needed to estimate frequencies we arrive at an effective maximum population size of 10^4-10^5 per experiment. Thus, even NGS permits only small dances around the local protein sequence-fitness space.

Purchasing thousands to millions of synthetically generated DNA sequences is still not an economically viable option for the average academic lab. Furthermore, established facile protocols for random mutagenesis like error-prone PCR [37] or chemical synthesis by doping [1] provide access only to a minority of possible codon substitutions, and there is often a large variance in the number of mutations introduced. Thus, robust methods for constructing large, user-defined DNA libraries are needed.

Generation of libraries with mutations at 1-4 defined positions have been demonstrated using homologous recombination and cassette mutagenesis. For applications such as lead candidate maturation the generation of comprehensive single-site saturation mutagenesis (CSM) libraries is desired. A CSM library contains all possible single amino acid substitutions at every position in the primary sequence. One could generate such libraries by performing separate saturation mutagenesis reactions for each position using QuikChange or similar methods. However, there are now three methods that can generate CSM libraries for gene-length targets with a single reaction: PALS [38], PFunkel [39], and Nicking Mutagenesis [40]. In PFunkel mutagenesis, single mutants are generated by thermocycling mutagenic oligos with template DNA at a low primer:template ratio in a single test-tube. While PFunkel has been demonstrated on multiple systems with excellent performance [28,30,36] the method requires a bacteriophage preparation of a Uracil-containing ssDNA template, which can be laborious. To overcome this, Wrenbeck et al. developed a similar method, Nicking Mutagenesis, which uses plasmid dsDNA as the reaction template [40].

1.1. DNA Read Length Restrictions

One major limitation of NGS is the inherent short read length (75 to 300 nucleotides for Illumina sequencing platform) (**Figure 3A**). As such, a mutation located outside of the read window would be invisible. Longer read lengths are possible using PacBio and Oxford Nanopore instruments but at the cost of reduced throughput and accuracy, respectively. Because of these limitations, many groups perform deep mutational scanning on small genes or on subsets of genes (tiling) (**Figure 3B**) [24,26,28,31,36,41].

An emerging strategy is to perform a selection on a full-length gene but ‘link’ or phase haplotypes from one portion of the gene to the remainder (**Figure 3C**) [38,42–48]. For example, Sarkisyan et al. introduced a random 20-nucleotide barcode at the C-terminal end of a library of green fluorescent protein variants whilst performing error-prone PCR [48]. Genotypes were barcode linked by sequencing both the N- and C- termini, with the N-terminus brought into proximity of the barcode with successive digestion and ligation reactions.

1.1. Sequencing Analysis

A crucial step in any NGS-utilizing experiment is to extract useful phenotypic data - binding, kinetics, thermodynamic stability, host organismal fitness, etc. - from raw sequencing reads. Many groups report site-specific preferences as an enrichment ratio. To that end, Fowler et al. developed Enrich, a python-based software that transforms raw sequencing counts from pre- and post-selection populations into per-allele enrichment ratios [49]. Similarly, Bloom developed a software that calculates enrichments using a likelihood-based treatment of mutation counts instead of simple ratios [50]. Woldring et al. developed ScaffoldSeq, a Python-based software for the analysis of partially diverse protein sequences for single site and pairwise amino acid frequencies across the population [51].

Normalization of these enrichment ratios to an unambiguous fitness metric like binding or catalytic efficiency is perhaps the least standardized portion of the deep mutational scanning pipeline and there is a need for a community-wide consensus on how to normalize. Kowalsky et al. describe a mathematical framework for normalizing enrichment ratios of variants assayed in deep mutational scanning experiments for FACS and growth-based selections [36]. Similar approaches are used for plate-based selections [30]. Finally, Abriata et al. developed a webserver, PsychoProt, for the analysis of functional data from saturation mutational libraries and protein sequence alignments for biophysical constraints using structural information [52].

1. Conclusion

NGS has been a transformative technology for many fields in the biological sciences, with protein science and engineering being no exception. Generation and analysis of fitness landscapes can inform on mechanisms of natural evolution and fundamentals of enzyme behavior. Notable advances in our ability to engineer affinity and specificity in protein-ligand interactions has been enabled by NGS, while enzyme and membrane protein engineering has lagged behind largely because of the lack of generalized HTS strategies. The utility of NGS enabled enzyme and membrane protein engineering awaits development of generalized HTS for these important classes of proteins. Accurate and facile sequencing of non-contiguous mutations (haplotyping), either through the use of barcoding or the advent of longer-read technologies, will improve and expand the utility of NGS protein engineering.

Acknowledgments

We thank D. Woldring for helpful discussions. This work was supported by a graduate research fellowship from the MSU PBHS Biotechnology Training Program, grant number NIH-T32-GM110523 (to E.E.W) and US National Science Foundation Career Award #1254238 CBET (to T.A.W.).

References

1. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S: **High-resolution mapping of protein sequence-function relationships**. *Nat. Methods* 2010, **7**:741–746.
2. Hietpas RT, Jensen JD, Bolon DNA: **Experimental illumination of a fitness landscape**. *Proc. Natl. Acad. Sci.* 2011, **108**:7896–7901.
3. Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, Gaskins G, Liu Z, Deming L, Axen SD, et al.: **Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting**. *Elife* 2016, **5**:e15802.
4. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat. Methods* 2014, **11**:801–807.
5. Boucher JI, Bolon DNA, Tawfik DS: **Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature.** *Protein Sci.* 2016, doi:10.1002/pro.2928.
6. Dekosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, et al.: **High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire**. *Nat. Biotechnol.* 2013, **31**:166–169.
7. Wang B, Kluwe CA, Lungu OI, Dekosky BJ, Kerr SA, Johnson EL, Tanno H, Lee C, Jung J, Rezigh AB, et al.: **Facile discovery of a diverse panel of anti-Ebola virus antibodies by immune repertoire mining.** *Sci. Rep.* 2015, **5**.
8. Wang B, Lee C, Johnson EL, Kluwe CA, Cunningham C, Tanno H, Crooks RM, Georgiou G, Ellington AD, Wang B, et al.: **Discovery of high affinity anti-ricin antibodies by B cell receptor sequencing and by yeast display of combinatorial VH: VL libraries from immunized animals.** *MAbs* 2016, doi:10.1080/19420862.2016.1190059.
9. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, Bradbury ARM: **Deep sequencing in library selection projects: What insight does it bring?** *Curr. Opin. Struct. Biol.* 2015, **33**:146–160.
10. Mahon CM, Lambert MA, Glanville J, Wade JM, Fennell BJ, Krebs MR, Armellino D, Yang S, Liu X, O'Sullivan CM, et al.: **Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential**. *J. Mol. Biol.* 2013, **425**:1712–1730.
11. •Woldring DR, Holec P V, Zhou H, Hackel BJ: **High-Throughput Ligand**

Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* 2015, **10**:e0138956.
An innovative use of NGS to design naïve protein binder libraries. NGS was used to uncover site-specific preferences for binders isolated from a fibronectin scaffold library. These preferences were hard-coded into a second, more robust library.

12. Strauch E-M, Fleishman SJ, Baker D: **Computational design of a pH-sensitive IgG binding protein**. *Proc. Natl. Acad. Sci.* 2014, **111**:675–680.
13. Procko E, Berguig GY, Shen BW, Song Y, Frayo S, Convertine AJ, Margineantu D, Booth G, Correia BE, Cheng Y, et al.: **A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells.** *Cell* 2014, **157**:1644–1656.
14. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers C a, Kamisetty H, Blair P, Wilson I a, et al.: **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing.** *Nat. Biotechnol.* 2012, **30**:543–8.
15. Koday MT, Nelson J, Chevalier A, Koday M, Kalinoski H, Stewart L, Carter L, Nieusma T, Lee PS, Ward AB, et al.: **A Computationally Designed Hemagglutinin Stem-Binding Protein Provides In Vivo Protection from Influenza Independent of a Host Immune Response**. *Plos Pathog.* 2016, **12**:e1005409.
16. Forsyth CM, Juan V, Akamatsu Y, Dubridge RB, Doan M, Ivanov A V, Ma Z, Polakoff D, Wilson K, Powers DB, et al.: **Deep mutational scanning of an antibody mammalian cell display and massively parallel Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing**. *MAbs* 2013, **5**.
17. Fujino Y, Fujita R, Wada K, Fujishige K, Kanamori T, Hunt L, Shimizu Y, Ueda T: **Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning**. *Biochem. Biophys. Res. Commun.* 2012, **428**:395–400.
18. ••Koenig P, Lee C V, Sanowar S, Wu P, Stinson J, Harris SF, Fuh G: **Deep Sequencing-guided Design of a High Affinity Dual Specificity Antibody to Target Two Angiogenic Factors in Neovascular Age-related Macular Degeneration**. *J. Biol. Chem.* 2015, **290**:21773–21786.
An excellent study combining NGS and structural information to design antibodies with potent affinity to two structurally unrelated conformational epitopes. A Fab with dual specificity for Ang2 and VEGF with high affinity for both was developed using DMS.
19. ••Jardine JG, Kulp DW, Havenar-Daughton C, Sarkar A, Briney B, Sok D, Sesterhenn F, Ereno-Orbea J, Kalyuzhniy O, Deresa I, et al.: **HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen**. *Science (80-.).* 2016, **351**:1458–1463.
A germline-targeting immunogen was engineered by exhaustive deep mutational scanning against dozens of germline-reverted and mature broadly neutralizing antibodies to HIV-1. This paper represents the most extensive NGS-aided protein engineering example to date.

20. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A: **Computational design of ligand-binding proteins with high affinity and selectivity.** *Nature* 2013, **501**:212–216.

21. Taylor ND, Garruss AS, Moretti R, Chan S, Arbing MA, Cascio D, Rogers JK, Isaacs FJ, Kosuri S, Baker D, et al.: **Engineering an allosteric transcription factor to respond to new ligands**. *Nat. Methods* 2016, **13**:177–183.

22. Doolan KM, Colby DW: **Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing.** *J. Mol. Biol.* 2015, **427**:328–340.

23. Van Blarcom T, Rossi A, Foletti D, Sundar P, Pitts S, Bee C, Melton Witt J, Melton Z, Hasa-Moreno A, Shaughnessy L, et al.: **Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing**. *J. Mol. Biol.* 2015, **427**:1513–1534.

24. Kowalsky CA, Faber MS, Nath A, Dann HE, Kelly VW, Liu L, Shanker P, Wagner EK, Maynard JA, Chan C, et al.: **Rapid Fine Conformational Epitope Mapping Using Comprehensive Mutagenesis and Deep Sequencing**. *J. Biol. Chem.* 2015, **290**:26457–26470.

25. •Schlinkmann KM, Honegger A, Türeci E, Robison KE, Lipovšek D, Plückthun A: **Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations.** *Proc. Natl. Acad. Sci.* 2012, **109**:9810–9815.
An elegant early demonstration of NGS for uncovering the contribution of specific residues to the stability of a bacterial-expressed GPCR.

26. Schlinkmann KM, Hillenbrand M, Rittner A, Künz M, Strohner R, Plückthun A: **Maximizing detergent stability and functional expression of a GPCR by exhaustive recombination and evolution.** *J. Mol. Biol.* 2012, **422**:414–428.

27. Egloff P, Hillenbrand M, Klenk C, Batyuk A, Heine P, Balada S, Schlinkmann KM, Scott DJ, Schütz M, Plückthun A: **Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli.** *Proc. Natl. Acad. Sci.* 2014, **111**:E655–E662.

28. Klesmith JR, Bacik J, Michalczyk R, Whitehead TA: **Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli**. *ACS Synth. Biol.* 2015, **4**:1235–1243.

29. Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs R a, Palzkill T: **Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution.** *J. Mol. Biol.* 2012, **424**:150–67.

30. Firnberg E, Labonte JW, Gray JJ, Ostermeier M: **A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape**. *Mol. Biol. Evol.* 2014, **31**:1581–1592.

31. Stiffler MA, Hekstra DR, Ranganathan R: **Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase**. *Cell* 2015, **160**:882–892.

32. van der Meer J-Y, Poddar H, Baas B, Miao Y, Rahimi M, Kunzendorf A, Merkerk R Van, Tepper PG, Geertsema EM, Thunnissen AWH, et al.: **Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective Michaelases**. *Nat. Commun.* 2016, **7**:1–16.

33. •Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS: **Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes**. *Nucleic Acids Res.* 2014, **42**:gku511.
Innovative study where multiple fitness landscape for an aminoglycoside antibiotic resistance gene were generated with difference substrates at varying concentrations. The group leverage the fact that landcapes can vary dramatically with selection conditions and used differential analysis to generate enzymes with orthogonal substrate specificities.

34. Steinberg B, Ostermeier M: **Environmental changes bridge evolutionary valleys**. *Sci. Adv.* 2016, **2**:e1500921.

35. Fowler DM, Stephany JJ, Fields S: **Measuring the activity of protein variants on a large scale using deep mutational scanning**. *Nat. Protoc.* 2014, **9**:2267–2284.

36. •Kowalsky CA, Klesmith JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA: **High-Resolution Sequence-Function Mapping of Full-Length Proteins**. *PLoS One* 2015, **10**:e0118193.
This paper details methodological considerations for performing deep mutational scanning of full-length proteins. This includes a universal sequence tiling method and analytical solutions to normalize data across independent FACS and growth-based selections.

37. Cirino PC, Mayer KM, Umeno D: **Generating mutant libraries using error-prone PCR.** In *Directed Evolution Library Creation: Methods and Protocols.* . Springer; 2003:3–9.

38. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J: **Massively parallel single-amino-acid mutagenesis**. *Nat. Methods* 2015, **12**:203–206.

39. Firnberg E, Ostermeier M: **PFunkel: Efficient, Expansive, User-Defined Mutagenesis**. *PLoS One* 2012, **7**:e52031.

40. •Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat. Methods* 2016, doi:10.1038/nmeth.4029.
Streamlined protocol for generating comprehensive single-site saturation mutagenesis libraries.

41. Melamed D, Young DL, Gamble CE, Miller CR, Fields S: **Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein**. *RNA* 2013, **19**:1537–1551.

42. Borgstrom E, Redin D, Lundin S, Berglund E, Andersson AF, Ahmadian A: **Phasing of single DNA molecules by massively parallel barcoding**. *Nat. Commun.* 2015, **6**.

43. Cho N, Hwang B, Yoon J, Park S, Lee J, Seo HN, Lee J, Huh S, Chung J, Bang D: **De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries.** *Nat. Commun.* 2015, **6**.

44. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J: **Parallel, tag-directed assembly of locally derived short sequence reads**. *Nat. Methods* 2010, **7**:119–122.

45. Hong LZ, Hong S, Wong HT, Aw PPK, Cheng Y, Wilm A, Sessions PF De, Lim SG, Nagarajan N, Hibberd ML, et al.: **BAsE-Seq : a method for obtaining long viral haplotypes from short sequence reads**. *Genome Biol.* 2014, **15**.

46. Kosuri S, Church GM: **Large-scale de novo DNA synthesis: technologies and applications**. *Nat. Methods* 2014, **11**:499–507.

47. Stapleton JA, Kim J, Hamilton JP, Wu M, Irber LC, Maddamsetti R, Briney B, Newton L, Burton DR, Brown TC, et al.: **Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing**. *PLoS One* 2016, **11**:e0147229.

48. •Sarkisyan KS, Bolotin DA, Meer M V, Usmanova DR, Mishin AS, Sharonov G V, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al.: **Local fitness landscape of the green fluorescent protein**. *Nature* 2016, **533**:397–401. ***The local fitness landscape of full-length GFP was analyzed using a barcoding strategy that enabled haplotyping of non-local mutations. The library comprised variants with up to four single nucleotide variations.***

49. Fowler DM, Araya CL, Gerard W, Fields S: **Enrich: Software for analysis of protein function by enrichment and depletion of variants**. *Bioinformatics* 2011, **27**:3430–3431.

50. Bloom JD: **Software for the analysis and visualization of deep mutational scanning data**. *BMC Bioinformatics* 2015, **16**:1–13.

51. Woldring DR, Holec P V, Hackel BJ: **ScaffoldSeq: Software for characterization of directed evolution populations.** *Proteins Struct. Funct. Bioinforma.* 2016, **84**:869–874.

52. Abriata LA, Bovigny C, Peraro MD: **Detection and sequence/structure mapping of biophysical constraints to protein variation in saturated mutational libraries and protein sequence alignments with a dedicated server**. *BMC Bioinformatics* 2016, **17**:1–13.

53. Thyme SB, Song Y, Brunette TJ, Szeto MD, Kusak L, Bradley P, Baker D: **Massively parallel determination and modeling of endonuclease substrate specificity**. 2014, **42**:13839–13852.

54. Romero PA, Tran TM, Abate AR: **Dissecting enzyme function with microfluidic-based deep mutational scanning**. *Proc. Natl. Acad. Sci.* 2015, **112**:7159–7164.

55. Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, Shendure J, Brzovic PS, Fields S, Klevit RE: **Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis**. *Proc. Natl. Acad. Sci.* 2013, **110**:E1263–E1272.

Figure 1 Overview of the steps involved in deep mutational scanning. A library of protein variants is generated. Often this is a comprehensive single-site saturation mutagenesis library. The library is subjected to a high-throughput selection or screen for function. Examples of commonly used selections and screens include survival or competitive growth-based selections, protein binding screens like phage or yeast surface display, and fluorescence reporter-based screens. Variants are quantified in the pre- and

post-selection populations with counting via deep sequencing. These pre- and post-selection counts are transformed to a normalized functional score and are used to generate fitness landscapes of the target protein.

Figure 2 Engineering of affinity and specificity in protein-ligand interactions using deep mutational scanning. A.) Consider a protein binder that recognizes two separate targets A and B. Deep mutational scanning is performed against each target in parallel. Site-specific preferences for the protein against each target are visualized by a heatmap. Mutations can be combined to impart binders with greater affinity to both targets (top panel, red box) or restrict specificity to a single target (bottom panel, blue box). In practice, mutations at multiple positions are combined to make a focused library that is subsequently screened. B.) The structural basis for specificity- and affinity- altering mutations identified by deep mutational scanning using a dual action Fab (green cartoon) to Ang2 (purple surface) and VEGF (orange surface) as an example [18]. Heavy Chain (HC) L93K can increase affinity to both targets presumably by increasing electrostatic complementarity. Here Ang2 and VEGF are colored by electrostatic surface potential and HC-L93 (green) and HC-K93 (pink) are shown as sticks. By contrast, HC F98I is strongly depleted for in the VEGF binding population most likely because of steric clashes. Structures were created using PyMol from the PDB IDs 4ZFG, 4ZFF.

Figure 3 Strategies to overcome read length limitations of NGS. A.) Mutations falling outside of a length ’readable’ by current sequencing technologies would be invisible. B.) In a gene tiling approach, mutational libraries are prepared such that mutations are

restricted to a stretch of DNA readable by NGS platforms. Parallel screens or selections for function are performed. C.) Molecular barcoding of library members provides a means to overcome NGS sequencing read length restrictions. Randomized DNA barcodes are assigned to library member (1). Variants and their corresponding barcodes are linked and cataloged (haplotyped) (2). After functional selection (3), variants in the pre- and post-selection populations are counted by sequencing barcodes (4).

Table 1 NGS-assisted studies of large enzyme libraries.

Gene	**Application**	**Selection employed**	**Reference**
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Deng et al.[29]
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Firnberg et al.[30]
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Stiffler et al.[31]
APH(3')II kinase	aminoglycoside antibiotic resistance	Growth competition	Melnikov et al.[33]
Homing endonucleases	Genome engineering	Survival	Thyme et al.[53]
Levoglucosan kinase	Biomass conversion	Metabolic growth	Klesmith et al.[28]
amiE aliphatic amidase	Multiple industrial	Metabolic growth	Wrenbeck et al. (unpublished)
Bgl3 β-glucosidase	Biomass conversion	Micro-fluidic	Romero et al.[54]
Ube4b E3 ubiquitin ligase	E3 ubiquitin ligase	Phage display	Starita et al.[55]