Deep mutational scanning: a new style of protein science

Douglas M Fowler[1] & Stanley Fields[1–3]

Mutagenesis provides insight into proteins, but only recently have assays that couple genotype to phenotype been used to assess the activities of as many as 1 million mutant versions of a protein in a single experiment. This approach—'deep mutational scanning'—yields large-scale data sets that can reveal intrinsic protein properties, protein behavior within cells and the consequences of human genetic variation. Deep mutational scanning is transforming the study of proteins, but many challenges must be tackled for it to fulfill its promise.

As the central players in the cell's machinery, proteins have been the subject of numerous mutagenesis approaches that seek to characterize their functions. Nonetheless, the ability to measure the effects of mutations in proteins has been limited to a relatively small number of mutations. But what if it were possible to know the functional consequences of every possible amino acid change at each position in a protein or the biochemical activity of hundreds of thousands of different protein variants, each containing two, three or even more mutations? Recent technologies known collectively as 'deep mutational scanning' make mutagenesis studies of this magnitude a reality.

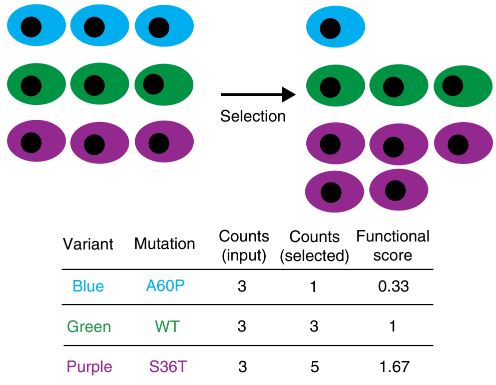
The key problem that deep mutational scanning solves is the limited ability to predict the most informative mutations in a protein to analyze. Changes to amino acids that are distant from binding or active sites can have drastic effects on the thermodynamic stability or enzymatic activity of a protein[1]. Highly conservative mutations, whose consequences can be difficult to predict, may be neutral, deleterious or hyperactivating[2,3]. Multiple mutations combined can lead to unexpectedly large increases or decreases in activity[4,5]. By enabling the impact of mutations to be examined in an unbiased fashion, deep mutational scanning can reveal the unexpected. It can also address otherwise intractable cases in which it is necessary to measure the activity of a huge number of variants. For example, functional analyses of genomes and of protein engineering experiments increasingly demand this scale of data.

Carrying out a deep mutational scan requires an assay amenable to a coupled genotype-phenotype platform (**Fig. 1**). Such platforms include cell-based assays, with a protein typically expressed from a plasmid or virus, or *in vitro* systems, such as phage or ribosome display. A library of mutated variants of the gene is synthesized, cloned into the appropriate vector and introduced, for example, into cells where the protein encoded by the gene carries out a function that can be selected for. The selection enriches cells with active protein variants and depletes those with inactive ones. The library is retrieved from both input and post-selection cells, and the frequency of each variant in the two libraries is determined by high-throughput DNA sequencing. The change in the frequency of each variant from input to selection serves as a measure of its function. Separation technologies, such as cell sorting, can also be used to place variants into bins, with the variants in each bin scored by DNA read counts.

The assays amenable to deep mutational scanning vary as widely as the activities that proteins can show. These include binding of a protein to a peptide, another protein, DNA, RNA or other ligands and enzymatic activities such as phosphorylation or ubiquitination. Cellular assays can take advantage of a growth or drug selection or expression of a protein that may be fluorescent or epitope tagged. *In vitro* approaches can enrich active variants on the basis of enzymatic activity, which can be combined with the use of an antibody that recognizes a post-translational modification. Because of the astronomical scale of DNA sequencing, millions of individual protein variants can be examined in a single experiment. This approach has been applied to a growing number of disparate proteins in a variety of contexts (**Table 1**). Establishing the infrastructure to carry out a deep mutational scan

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Medicine, University of Washington, Seattle, Washington, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to D.F. (dfowler@uw.edu) or S.F. (fields@uw.edu).

RECEIVED 7 JUNE 2013; ACCEPTED 19 MAY 2014; PUBLISHED ONLINE 30 JULY 2014; DOI:10.1038/NMETH.3027



Variant	Mutation	Counts (input)	Counts (selected)	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67

Figure 1 | Deep mutational scanning generates large-scale mutational data. Deep mutational scanning draws on high-throughput DNA sequencing to assess the functional capacity of a large number of protein variants simultaneously. A library of protein variants is created and introduced into a system in which the genotype of each variant is linked to a selectable phenotype, then selection for the function of the protein is imposed (top). Variants with high activity increase in frequency, whereas variants with low activity decrease in frequency. High-throughput DNA sequencing is used to measure the frequency of each variant before and after selection. These frequency data are analyzed to generate a functional score for each protein variant (bottom). WT, wild type.

for the first time can be challenging, but it is becoming less so as reagents, software and methods are developed (see ref. 6 for protocols).

On the simplest level, the large-scale mutational data that result from a deep mutational scan reveal the functional consequences of all possible single mutations. These data can be organized into a sequence-function map (**Fig. 2**). Such a map can be viewed as an all-residue scan, in which each position has been mutated to every other amino acid. These maps are dense with information, with each position having a unique pattern of functional effects; most substitutions are likely to be deleterious, but a few may enhance activity. In addition to characterizing the effects of single mutations, deep mutational scanning can also examine the effects of multiple mutations. Collectively, these data can yield insights into protein structure and function, but gleaning these insights is a challenge for both experimental and computational biologists.

Inference of fundamental protein properties

A number of biochemical methods are customarily used to directly assay the fundamental properties of proteins; for example, chemical denaturation analyzes thermodynamic stability, enzyme kinetics reveal mechanism, X-ray crystallography provides structure, and light scattering measures particle size. These methods apply purpose-built instrumentation in the context of a specialized workflow, generally feasible for no more than a handful of variants.

Instead of using such methods to measure protein properties in a serial fashion, one might infer some properties from large-scale mutational data. This approach draws on knowledge derived from more than a century of study of proteins, including principles of how they fold and unfold, act in catalysis, interact with solvents and evolve. For a given protein property, such prior knowledge has the potential to generate a model or algorithm that relates the functional consequences of mutations to the property in question, and the model could be applied to the large-scale data obtained in a deep mutational scan (**Box 1**). This approach could augment and eventually supplant some traditional methods that are time-, cost- and labor-intensive. We highlight three areas in which this approach has progressed.

First, because stabilizing mutations can rescue destabilizing ones[7–9], large-scale mutational data can be analyzed to identify thermodynamically stabilizing mutations. These mutations are important for engineering proteins for pharmaceutical or industrial uses and are difficult to identify; most mutations are either neutral or destabilizing. Current methods to identify stabilizing mutations have limitations, including poor performance for large or atypical proteins, extensive validation requirements, limited output and the identification of mutations that, although stabilizing, also result in an unintended loss of activity[10–12]. In previous work, we developed a computational model that measures the effectiveness of single mutations in rescuing many other deleterious single mutations when they occur together in a doubly mutated variant[13]. We applied this model to measurements of the peptide-binding capacities of ~50,000 variants of a WW domain and identified new stabilizing mutations.

Second, because mutations can perturb enzyme function, analysis of large-scale mutational data can reveal aspects of a protein's catalytic mechanism. It is possible to identify rare variants that have enhanced activity or altered specificity. Such unusual variants were recently identified on the basis of the ubiquitination activity of ~100,000 variants of an E3 ubiquitin ligase, and these hyperactive variants were used to unlock mechanistic details through further biochemical and structural approaches[14]. Other analyses of enzyme mechanism from large-scale mutagenesis data

Table 1 | Deep mutational scanning targets

Scanned protein	Model	Selection
Fab antibody fragment[36]	Ribosome display	Ligand binding
YAP65 WW domain[13,37]	T7 bacteriophage	Ligand binding
E4B ubiquitin ligase[14]	T7 bacteriophage	Ubiquitination activity
PKA regulatory subunit[38]	T7 bacteriophage	Ligand binding
Synthetic PDZ domain[39]	M13 bacteriophage	Ligand binding
CcdB[16]	*Escherichia coli*	Toxin activity
PSD95 PDZ domain[40]	*E. coli*	Ligand binding
G protein–coupled receptor[41]	*E. coli*	Ligand binding
Designed influenza inhibitor[29]	*Saccharomyces cerevisiae* surface display	Ligand binding
Designed lysozyme inhibitor[42]	*S. cerevisiae* surface display	Ligand binding
Designed digoxigenin binder[43]	*S. cerevisiae* surface display	Small molecule binding
IgG1 CH3 domain[44]	*S. cerevisiae* surface display	Ligand binding after thermal stress
Hsp90 (refs. 45,46)	*S. cerevisiae* complementation	Growth rate
Matα2 degron[23]	*S. cerevisiae* fusion protein	Growth rate
Ubiquitin[47]	*S. cerevisiae* complementation	Growth rate
Pab1 (ref. 17)	*S. cerevisiae* complementation	Growth rate
Neuraminidase[48]	Mammalian cell	Oseltamivir resistance
IgG CDRs[49]	Mammalian cell display	Ligand binding
B-Raf[50]	Mammalian cell	Vemurafenib resistance

Fab, fragment antigen binding; YAP65, yes-associated protein-65; PKA, protein kinase A; PSD95, postsynaptic density protein-95; Hsp90, heat-shock protein 90; Matα2, mating type protein-α2; Pab1, poly(A)-binding protein-1; CDR, complementarity-determining region.

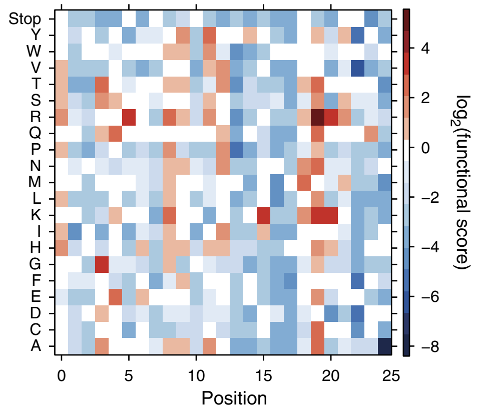


Figure 2 | Large-scale mutational data illustrate how protein sequence affects function. A hypothetical sequence-function heat map is shown for a 25-residue portion of a protein, illustrating the functional consequences of making every single amino acid mutation at every position. Positions are indicated numerically, and each mutation is indicated by its single-letter amino acid abbreviation. The color of each element of the heat map

make use of the observation that the most mutation-intolerant positions in a protein frequently correspond to residues directly involved in contacting the substrate or performing catalysis. Another potential starting point is the pairing of hyperactivating single mutations with deleterious single mutations that affect folding, stability, substrate interaction or other properties. For example, a mutation that enhances catalysis might be expected to rescue deleterious mutations that destabilize the protein but not those that block substrate binding.

Third, because mutations can perturb protein structure, large-scale mutational data can contribute to structural efforts. X-ray crystallography and nuclear magnetic resonance yield detailed structures, but do not work for every protein, particularly transmembrane proteins and large protein complexes[15]. *De novo* prediction of protein structure, although useful, cannot routinely provide useful structures even of average-sized proteins. Mutational data can help discriminate among predicted protein structures. For example, the functional consequences of mutation at each position in the bacterial toxin CcdB correlate with distance to the protein surface in a known structure. Adkar *et al.*[16] used this observation to select accurate CcdB structures from among a large set of predictions on the basis of which positions were buried. In another example from our own work, positions found to be sensitive to most substitutions except those to hydrophobic amino acids constituted the core hydrophobic structure of the protein[17].

In the future, large-scale mutational data could facilitate the prediction of protein secondary structure. Typically, algorithms base predictions on the amino acid preferences in each type of secondary structure (α-helix, β-sheet or loop) in a training set of proteins with known structures[18,19]. As an alternative, large-scale mutational data on proteins with known structures could also reveal amino acid preferences within structural elements, and the resulting preferences could be used to enhance structure prediction algorithms. A provocative challenge is the use of deep mutational scanning data to generate structural models. We suggest that these data could be analyzed to determine covarying positions in a protein's sequence, with the expectation that these positions will be close to one another in the three-dimensional folded structure. These experimentally determined distance constraints could then be combined with protein structure modeling software such as Rosetta to produce a plausible structural model[20]. Indeed, the fact that covariation between positions derived from the natural evolution of a protein can be used to predict structure if the multiple sequence alignment for the protein is sufficiently large[21] hints that this approach is feasible.

Analyzing large-scale mutational data is challenging because the principles by which fundamental protein properties relate to mutational data are not fully understood (**Box 1**). In some cases, lessons learned from the study of a small number of mutations will generalize well, but in others some refinement of our understanding will be required. Furthermore, these analyses require high-quality mutational data to succeed. High-throughput methods are notoriously susceptible to problems with data quality. Thus, practitioners will need to develop and apply standards, especially regarding appropriate replication and models for controlling systematic and stochastic error. Nevertheless, there is a huge potential payoff: a common method for understanding fundamental properties of proteins in their native environment.

Understanding how proteins behave in cells

Deep mutational scanning can be conducted in cells and thus offers the opportunity to marry protein science with cell-based approaches. Furthermore, the power of the technology is magnified by the fact that, for a particular protein, scans can be redone in a number of 'sensitized' backgrounds or conditions (**Fig. 3**)—a veritable 'Hershey heaven'[22], where repeating the same experiment with slight alterations yields novel data. We discuss three examples of this approach.

First, deep mutational scanning can be used to probe protein-protein interactions. Structural approaches for studying such interactions, such as cocrystallization, yield high-resolution information, but their throughput is inherently low. High-throughput approaches, such as yeast two-hybrid or mass spectrometry, provide little, if any, structural detail. A library of variants can be screened for interaction in cells that overproduce a partner protein. The expectation is that a subset of mutations that in the initial (nonsensitized) screen were deleterious might be neutral in the presence of excess binding partner, revealing positions in the protein relevant to the interaction.

Second, mutational scanning can measure the stability in cells of protein variants that are tagged with a required metabolic enzyme[23]. If the stability of the enzyme depends on the stability of the variant to which it is fused, then cells harboring a long-lived variant will have high concentrations of the enzyme and grow faster. The influence of protein-degradation factors could be investigated by varying their abundance.

Third, mutational scanning using cell-based protein-aggregation models could yield details of the biophysical processes driving aggregation *in vivo*. For example, variants of an aggregation-prone protein could be fused to an essential enzyme whose activity diminishes as the aggregation state of the variant increases[24]. Furthermore, by again varying the expression of chaperones and degradation factors, the experimenter might better understand how these factors identify and degrade aggregation-prone proteins.

BOX 1 INTERPRETING LARGE-SCALE MUTATIONAL DATA

The initial stages of data analysis focus on producing a set of high-quality functional scores from raw sequence data[51]. In the simplest case, reads are aligned to a wild-type template, variants are enumerated and functional scores are calculated by taking the ratio of the frequency of each variant before and after selection[37]. More complex cases (for example, those incorporating time-series data) can be dealt with using linear models[13,14]. Nevertheless, clear standards for analyzing deep mutational scanning data have yet to emerge. Enrich, an interactive software package for accomplishing the first data-analysis phase, is publicly available, but its use requires command-line expertise[52]. Enrich guides users through the process of transforming raw high-throughput sequencing data into a set of variant functional scores. Enrich also generates a comprehensive sequence-function map from the data. However, deeper analyses of the functional scores are considerably more challenging and depend on the questions being asked. In some cases, analytical paradigms are already emerging, including those that examine how multiple mutations interact and how large-scale mutagenesis data change under different experimental conditions. Data analysis remains a significant challenge, but not an intractable one.

For example, when engineering a protein or when classifying mutations in a disease-related protein, the experimenter may be interested only in how single mutations affect protein activity. In this case, data for single amino acid substitutions derived from a deep mutational scan can be displayed as a heat map relating sequence to function (**Fig. 2**). Further analysis can yield insights into topics such as fundamental protein properties, the behavior of proteins inside cells and the paths of protein evolution but is typically a slow and complex undertaking.

Successful interpretation of deep mutational scanning data starts with proper experimental design. Will the experimenter take advantage of direct selection for a protein property of interest? Will the analysis require only single mutations, or will multiple mutations be needed? Will the analysis need large numbers of variants, or will a few thousand suffice? To give an idea of how one might answer these questions, we highlight three broad experimental designs and give examples of how an experimenter might go about analyzing the resulting data sets.

Direct selection for a protein property of interest results in the most straightforward analysis of large-scale mutational data. Examples include measurement of:

- thermodynamic stability of a library of IgG variants, using yeast display selection and thermal denaturation[44];
- *in vivo* protein stability of a library of yeast degron variants, using a metabolic reporter protein fusion[23]; and
- inhibitor resistance of a library of BRAF variants, using a cell-based resistance assay[50].

Knowledge-based inference is a more complex type of analysis and can be applied when direct selection is not possible for the desired protein property. For example, directly selecting for mutations that change an enzyme's mechanism would be difficult. Here the experimenter selects for protein function without using specialized conditions (for example, higher temperature to select for stability or the presence of an inhibitor to select for resistance) and then carries out an analysis that relates the functional scores to the property of interest. Examples include identification of:

- thermodynamically stabilizing mutations, identified because they rescue multiple destabilizing mutations[13];
- buried positions, identified because they tolerate fewer substitutions than solvent-exposed ones[16];
- core positions, identified because they exhibit similar patterns of preference for hydrophobic amino acids[17]; and
- mechanism-altering mutations, identified because they are hyperactivating[14].

In even more complex cases, no analytic framework for the mutational data yet exists and will need to be developed. Examples include:

- benchmarking and improving computational approaches for interpreting human genetic variation;
- improving the correlation of biochemical properties with disease risk;
- enhancing prediction algorithms for *de novo* protein structure and activity; and
- understanding protein evolution.

Protein evolution and engineering

Experimental evolution approaches offer the opportunity to watch protein evolution as it occurs, but they have been limited either to examining a handful of variants or to making population-based measurements. Owing to the vastness of the sequence landscape, conclusions arising from these studies have been incomplete and sometimes contradictory. Protein evolution has also been treated theoretically, but many predictions remain untested. Deep mutational scanning approaches, when applied to experimental evolution of proteins, offer the ability to explicitly and simultaneously track the fates of hundreds of thousands of sequences. They can thus begin to address fundamental questions[25], such as: how many paths can evolution take? How many mutations are required to produce a new function? Are there many distinct sequences that could evolve to solve the same problem? In short, these approaches offer the opportunity to experimentally explore the protein fitness landscapes that shape evolutionary trajectories. For example, large-scale mutational data on a WW domain[13] and on an HIV protease and reverse transcriptase[26] have revealed that some combinations of mutations within variants interact to produce unexpectedly large functional effects, such that intramolecular mutation interaction 'hotspots' can be identified within these

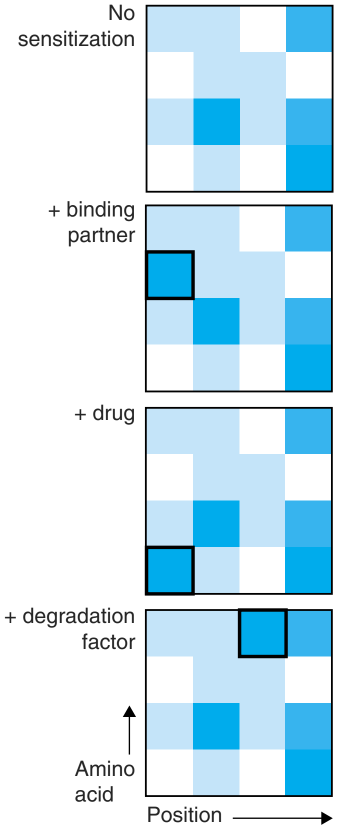


Figure 3 | Deep mutational scanning in sensitized backgrounds as a strategy for uncovering protein features. Hypothetical sequence-function heat maps collected under different conditions are shown. Once a deep mutational scan has been done, it can be repeated in a sensitized background, which can be created by altering the cellular or chemical environment in which the scan is conducted. The difference in functional effect for a particular mutation in a sensitized background could reveal the importance of an amino acid at a given position for the process under study.

proteins. High-throughput sequencing of T7 RNA polymerase evolving to bind new promoter sequences has revealed distinct classes of convergently evolved solutions[27].

Deep mutational scanning experiments should also be instrumental in realizing the promise of protein engineering, which improves existing proteins, and *de novo* design, which imagines new ones with desired features. Currently, efforts in these areas proceed from rule-based design[28] or use blind selection to identify one or a few variants with improved functionality among a library. In both cases, deep mutational scanning approaches could be transformative, enabling the identification of large numbers of useful mutations that can be combined to refine engineered or designed proteins. For example, this approach was used to optimize a computationally designed hemagglutinin-binding protein that inhibits influenza virus[29], resulting in the identification of five mutations that combined to produce a 25-fold improvement in affinity. Traditional affinity-maturation approaches would not have resulted in the final, high-affinity inhibitor because such approaches cannot effectively explore the staggeringly large number of mutant combinations required to find a variant with five mutations. Large-scale mutagenesis data offer the opportunity to improve the protein-design process by enabling designers to examine exhaustively where and why their algorithms fail[29,30].

Deep mutational scanning and human genetics

A large component of the genetic basis of disease lies in rare variation, with every human carrying, on average, ~300 rare, protein-encoding variants[31]. Knowing the functional consequences of rare mutations in important genes is crucial for many people, from physicians, pharmacists and patients to casual users of personalized DNA testing. Most existing experimental approaches are not practical for assessing the rapidly increasing number of these rare mutations being identified. They simply cannot achieve the scale necessary to measure the phenotypic consequences of the variation that can occur in a typical human protein, which comprises 375 amino acids subject to 7,500 possible single mutations (including to stop codons)[32]. The challenge is highlighted by the fact that 10% of women harboring a missense mutation in the *BRCA1* gene, which may predispose them to breast cancer, are told they harbor a "variant of unknown significance"[33]. That BRCA1, one of the best-studied proteins, still generates such diagnoses indicates that the situation for the average protein implicated in human disease is far worse. Furthermore, it will not be possible to repeat the investment of time and money in studying BRCA1 for each of these thousands of other proteins.

Currently, computational prediction of the functional consequences of mutations with programs such as Condel, GERP, PolyPhen-2 and SIFT is the best researchers can do. But these computational approaches are limited in their accuracy[34]. For example, when Condel, PolyPhen-2 and SIFT predict the functional consequences of a set of known deleterious mutations, they produce correct and concordant results in fewer than half the cases[35]. Because these tools are based on evolutionary conservation of individual positions and/or the physicochemical properties of amino acids, they are relatively successful only on average. But they fail in an unacceptably large fraction of cases, making them far from ideal for clinical use.

Large-scale mutational data could empower these computational approaches. First, these data provide a new resource for benchmarking computational approaches. Second, analysis of a modest number of large-scale mutagenesis data sets derived from proteins with diverse structures and functions could enhance our understanding of how, in a general sense, mutations affect protein function. This information should be useful for improving the accuracy of physicochemical models of the impact of mutations. Third, large-scale mutagenesis data in model organisms that are selected for their fitness could even contribute to developing computational models that predict the effects of mutations on a more complex organism.

In principle, experimental characterization of the functional consequences of all possible single amino acid substitutions using a deep mutational scanning approach could obviate the need for computational inference in interpreting coding variation by furnishing sequence-function maps of disease-related proteins (**Fig. 2**). This task seems daunting, as it would require thousands of sequence-function maps for proteins with an enormous range of functions. However, the challenge may not be quite as formidable as it seems: many disease-related proteins fall into well-studied classes, such as transcription factors, protein kinases, surface receptors and DNA-repair proteins, which may allow the use of some existing, generic assays (**Fig. 4**). Of course, before such data are applied in the clinic, assays to determine protein function scores must be vetted for their capacity to reflect disease risk,

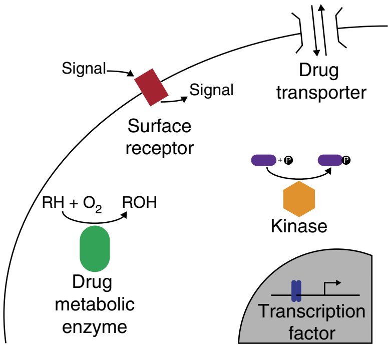


Figure 4 | Sequence-function maps of proteins important in disease. A schematic of a hypothetical cancer cell is shown; mutations in drug transporters, drug metabolic enzymes, transcription factors and signaling proteins all have the capacity to influence the effectiveness of treatment. Deep mutational scanning of cancer-related proteins could revolutionize our understanding of the consequences of mutations in these proteins and enable genomic medicine.

pathogenicity or progression. The jury is still out on which *in vitro* assays will do so adequately. Furthermore, a simple functional assay that is amenable to a deep mutational scan cannot be generated for every protein. The possible rewards for such an approach are nevertheless considerable. A large, coordinated project could, for example, generate sequence-function maps for a set of cancer-related proteins, providing an invaluable clinical resource.

Unresolved questions

Between the promise and the reality of deep mutational scanning lie many questions. Is there as much useful protein information latent within these large data sets as we speculate? It is clear already that large-scale mutational data contain a rich array of information. But developing analytic methods to reveal some of this information, such as protein structure, will probably require substantial development. Furthermore, it may be difficult to design assays that couple some cell-based properties, such as localization or post-translational modification, to the sequencing readout required for a deep mutational scan.

For a scan to be effective, the development of an appropriate assay for the function of interest is perhaps even more important than the methods used for mutagenesis, library construction, sequencing and computational analysis. Can the scale of assay development match the pace of progress in DNA synthesis and sequencing? It is crucial that the selection condition alter or separate library members according to their functional capacity, ideally across a wide range of activity levels. The assay must enable the production of DNA libraries that are amenable to high-throughput sequencing, something not every assay does. Although researchers can draw on decades of collective experience in crafting these functional assays, choosing and calibrating one that works at high throughput remains a formidable undertaking.

Will these approaches be put into place soon enough to deal with the deluge of human genetic variation now being discovered, and will the mutational data generated *in vitro* adequately reflect the complex roles of disease proteins? The concern is that simple assays that can be scored at high throughput may not adequately reflect human disease. The limits of simple assays must undoubtedly be respected. For example, assays for proteins that act extracellularly or that are poorly conserved are probably not good candidates. Assays for well-conserved intracellular proteins will probably be useful, but they will need to be validated to ensure they adequately reflect disease risk. Advances in genome editing could pave the way for deep mutational scanning experiments in human cell lines, partially alleviating this concern, although even human cell–based assays are limited in their ability to model organ or whole-organism disease phenotypes. We suggest that for proteins with simple molecular functions (for example, metabolic enzymes), large-scale mutagenesis data might have potential for direct use in the clinic. For proteins with complex functions (for example, signaling proteins), large-scale mutagenesis data will need to be combined with an integrative computational model. In either case, extensive sets of protein variants whose activity scores can be compared to known disease risk and outcome will be needed to establish clinical utility of these data.

In summary, deep mutational scanning can be used to generate large-scale mutational data for nearly any protein. Because this approach is rooted in a rapidly developing technology—high-throughput sequencing—it is likely that its power and scope will continue to grow. We have highlighted some of the ways in which large-scale mutational data could transform protein science. The many challenges to this transformation also provide many opportunities to protein scientists. Understanding the vast number of protein variants in humans demands that experimental and computational methods be developed. Deep mutational scanning strategies provide one avenue to address this need.

ACKNOWLEDGMENTS
We thank A. Merz, M. Hochstrasser, C. Queitsch, A. Gitler, J. Bloom, E. Marcotte, E. Phizicky and M. Wickens for helpful discussions and comments. This work was supported by P41 GM103533 (to S.F.) and F32 GM084699 (to D.M.F.) from the US National Institute of General Medical Sciences. S.F. is supported by the Howard Hughes Medical Institute.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Freeman, A.M., Mole, B.M., Silversmith, R.E. & Bourret, R.B. Action at a distance: amino acid substitutions that affect binding of the phosphorylated CheY response regulator and catalysis of dephosphorylation can be far from the CheZ phosphatase active site. *J. Bacteriol.* **193**, 4709–4718 (2011).
2. Jonson, P.H. & Petersen, S.B. A critical view on conservative mutations. *Protein Eng.* **14**, 397–402 (2001).
3. Gilbert, G.E., Novakovic, V.A., Kaufman, R.J., Miao, H. & Pipe, S.W. Conservative mutations in the C2 domains of factor VIII and factor V alter phospholipid binding and cofactor activity. *Blood* **120**, 1923–1932 (2012).
4. Zhang, W., Dourado, D.F.A.R., Fernandes, P.A., Ramos, M.J. & Mannervik, B. Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochem. J.* **445**, 39–46 (2012).
5. Natarajan, C. *et al.* Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324–1327 (2013).
6. Fowler, D.M., Stephany, J.J. & Fields, S. Measuring the activity of protein variants on a large-scale using deep mutational scanning. *Nat. Protoc.* doi:10.1038/nprot.2014.153 (in the press).
7. Wang, X., Minasov, G. & Shoichet, B.K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).
8. Bloom, J.D. & Arnold, F.H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. USA* **106**, 9995–10000 (2009).
9. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D.S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).

10. Potapov, V., Cohen, M. & Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.* **22**, 553–560 (2009).
11. Magliery, T.J., Lavinder, J.J. & Sullivan, B.J. Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Curr. Opin. Chem. Biol.* **15**, 443–451 (2011).
12. Foit, L. *et al.* Optimizing protein stability *in vivo*. *Mol. Cell* **36**, 861–871 (2009).
13. Araya, C.L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA* **109**, 16858–16863 (2012).
14. Starita, L.M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. USA* **110**, E1263–E1272 (2013).
15. Lander, G.C., Saibil, H.R. & Nogales, E. Go hybrid: EM, crystallography, and beyond. *Curr. Opin. Struct. Biol.* **22**, 627–635 (2012).
16. Adkar, B.V. *et al.* Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* **20**, 371–381 (2012).
17. Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
18. Aydin, Z., Singh, A., Bilmes, J. & Noble, W.S. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. *BMC Bioinformatics* **12**, 154 (2011).
19. Chen, K. & Kurgan, L. Computational prediction of secondary and supersecondary structures. *Methods Mol. Biol.* **932**, 63–86 (2013).
20. Kim, D.E., DiMaio, F., Yu-Ruei Wang, R., Song, Y. & Baker, D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* **82**, 208–218 (2014).
21. Marks, D.S., Hopf, T.A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
22. Creager, A.N.H. Hershey heaven. *Nat. Struct. Biol.* **8**, 18–19 (2001).
23. Kim, I., Miller, C.R., Young, D.L. & Fields, S. High-throughput analysis of *in vivo* protein stability. *Mol. Cell. Proteomics* **12**, 3370–3378 (2013).
24. Morell, M., de Groot, N.S., Vendrell, J., Avilés, F.X. & Ventura, S. Linking amyloid protein aggregation and yeast survival. *Mol. Biosyst.* **7**, 1121–1128 (2011).
25. Dean, A.M. & Thornton, J.W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat. Rev. Genet.* **8**, 675–688 (2007).
26. Hinkley, T. *et al.* A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.* **43**, 487–489 (2011).
27. Dickinson, B.C., Leconte, A.M., Allen, B., Esvelt, K.M. & Liu, D.R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc. Natl. Acad. Sci. USA* **110**, 9007–9012 (2013).
28. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
29. Whitehead, T.A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).
30. Moretti, R. *et al.* Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins* **81**, 1980–1987 (2013).
31. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
32. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400 (2005).
33. Millot, G.A. *et al.* A guide for functional analysis of *BRCA1* variants of uncertain significance. *Hum. Mutat.* **33**, 1526–1537 (2012).
34. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14**, S7 (2013).
35. Gray, V.E., Kukurba, K.R. & Kumar, S. Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics* **28**, 2093–2096 (2012).
36. Fujino, Y. *et al.* Robust *in vitro* affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochem. Biophys. Res. Commun.* **428**, 395–400 (2012).
37. Fowler, D.M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
38. Gold, M.G. *et al.* Molecular basis of AKAP specificity for PKA regulatory subunits. *Mol. Cell* **24**, 383–395 (2006).
39. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* **6**, 1782–1790 (2010).
40. McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).
41. Schlinkmann, K.M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein–coupled receptor uncovered by all-versus-all mutations. *Proc. Natl. Acad. Sci. USA* **109**, 9810–9815 (2012).
42. Procko, E. *et al.* Computational design of a protein-based enzyme inhibitor. *J. Mol. Biol.* **425**, 3563–3575 (2013).
43. Tinberg, C.E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
44. Traxlmayr, M.W. *et al.* Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *J. Mol. Biol.* **423**, 397–412 (2012).
45. Jiang, L., Mishra, P., Hietpas, R.T., Zeldovich, K.B. & Bolon, D.N.A. Latent effects of Hsp90 mutants revealed at reduced expression levels. *PLoS Genet.* **9**, e1003600 (2013).
46. Hietpas, R.T., Jensen, J.D. & Bolon, D.N.A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. USA* **108**, 7896–7901 (2011).
47. Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D. & Bolon, D.N.A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425**, 1363–1377 (2013).
48. Wu, N.C. *et al.* Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. *J. Virol.* **87**, 1193–1199 (2013).
49. Forsyth, C.M. *et al.* Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs* **5**, 523–532 (2013).
50. Wagenaar, T.R. *et al.* Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res.* **27**, 124–133 (2014).
51. Araya, C.L. & Fowler, D.M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* **29**, 435–442 (2011).
52. Fowler, D.M., Araya, C.L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).