Viewing Protein Fitness Landscapes Through a Next-Gen Lens

Jeffrey I. Boucher, Pamela Cote, Julia Flynn, Li Jiang, Aneth Laban, Parul Mishra, Benjamin P. Roscoe, and Daniel N. A. Bolon¹

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605

ORCID ID: 0000-0001-5857-6676 (D.B.)

ABSTRACT High-throughput sequencing has enabled many powerful approaches in biological research. Here, we review sequencing approaches to measure frequency changes within engineered mutational libraries subject to selection. These analyses can provide direct estimates of biochemical and fitness effects for all individual mutations across entire genes (and likely compact genomes in the near future) in genetically tractable systems such as microbes, viruses, and mammalian cells. The effects of mutations on experimental fitness can be assessed using sequencing to monitor time-dependent changes in mutant frequency during bulk competitions. The impact of mutations on biochemical functions can be determined using reporters or other means of separating variants based on individual activities (e.g., binding affinity for a partner molecule can be interrogated using surface display of libraries of mutant proteins and isolation of bound and unbound populations). The comprehensive investigation of mutant effects on both biochemical function and experimental fitness provide promising new avenues to investigate the connections between biochemistry, cell physiology, and evolution. We summarize recent findings from systematic mutational analyses; describe how they relate to a field rich in both theory and experimentation; and highlight how they may contribute to ongoing and future research into protein structure–function relationships, systems-level descriptions of cell physiology, and population-genetic inferences on the relative contributions of selection and drift.

ADVANCED nucleic acid sequencing technologies (Margulies *et al.* 2005; Bentley *et al.* 2008; Eid *et al.* 2009; Rothberg *et al.* 2011) have transformed biological research in manners that were commonly anticipated (e.g., delineating genetic blueprints for different individuals or species) and in largely unexpected fashions that have come from widespread access to high-throughput DNA sequencing. The commoditization of massively parallel sequencing has made it accessible to many researchers (e.g., it currently costs about \$1000 to obtain >20 million reads of 100 bases). Due to favorable cost efficiency and data quality, modern sequencing approaches have become appealing for estimating the relative abundance of nucleic acid molecules in complex mixtures. For example, sequencing approaches are currently utilized to monitor the abundance of RNA molecules in cells (Nagalakshmi *et al.* 2008; Wang *et al.* 2009) and in subcellular locations (Ingolia

et al. 2009). These pioneering studies demonstrated the capabilities of sequencing to quantify the relative abundance of thousands of different nucleic acid molecules. Building on this concept, a growing number of researchers have started using sequencing approaches to analyze the frequency of thousands of individual variants in libraries of engineered mutations. Monitoring the frequency change of engineered variants in response to selection pressures provides insights into landscapes of mutations and their impacts on function and/or experimental fitness (Fowler *et al.* 2010; Hietpas *et al.* 2011; McLaughlin *et al.* 2012).

A Perspective on Mutational Landscapes

Landscapes of mutations in the broadest sense include all possible mutational combinations, an almost infinite complexity that remains inaccessible to experimental approaches. The concept of a mutational landscape dates back to the first half of the 20th century, prior to the discovery of the genetic code and even the recognition of DNA as the genetic material, and was the result of visionary inferences by Wright (1932). The results of genetic crosses at this time indicated that there were

~1000 genes in an organism, leading Wright to speculate that, “with 10 allelomorphs in each of 1000 loci, the number of possible combinations is 10^{1000} , which is a very large number. It has been estimated that the total number of electrons and protons in the visible universe is much less than 10^{100} .” While mindful of the immensity of combinatorial allele space, Wright went on to consider how individuals and populations could sample this space during natural selection. This led Wright to diagram a landscape of combinatorial allele space where field lines indicated adaptive fitness, akin to the depiction of elevation in topographical maps (Figure 1A). As Wright noted, this landscape view of adaptation included several simplifications. For example, the dimensional representation of allele space was vastly simplified (from 1000-fold to 2-fold) and fitness was represented as a continuous surface, suggesting that adjacent positions in allele space exhibit similar fitness. Despite these simplifications, the adaptive landscape provided a compelling framework where Wright considered, “a mechanism by which the species may continually find its way from lower to higher peaks,” including the impacts of mutation rate, selection strength, environmental changes, and population demography.

The adaptive mechanisms outlined by Wright remain a focus of modern experiments seeking to understand molecular or mutational pathways of adaptation. Many of these studies aim to map out the landscape of potential mutational steps required to convert an ancestral gene into a derived gene with a biochemically defined adaptive advantage (Figure 1B). While this type of study focuses on combinations of mutations within the same gene that are distinct from the allele space considered by Wright, many of the same principles and questions apply. By what mechanism(s) can the ancestral sequence convert to the derived sequence with increased fitness? Are there many pathways available, or does the shape of the landscape impose a barrier to certain pathways?

Experimental studies have begun to answer these questions for a handful of systems. In the β -lactamase protein, analyses of all 32 possible mutational combinations between a drug-sensitive and a drug-resistant variant (differing at five amino acid positions) indicated that the fitness landscape imposed barriers to many of these 32 possible adaptive pathways (Weinreich *et al.* 2006). As exemplified by this work, identifying ancestral sequences is valuable to understanding evolutionary processes. The evolutionary relationship for many genes can now be inferred thanks to the explosion of available DNA sequences from extant species and the development of maximum likelihood approaches (Yang 1997). As pioneered by Joe Thornton, powerful phylogenetic tools can be utilized in reconstructing and analyzing ancestral proteins to understand molecular mechanisms, including the evolution of novel functions (Thornton *et al.* 2003). Ancestral protein reconstruction studies have highlighted an important role for permissive mutations that do not impact function directly, but instead enable additional mutations to alter function as in the evolution of new substrate recognition in nuclear steroid receptors (Ortlund *et al.* 2007). These impressive studies of

β -lactamase and steroid receptors demonstrate the value of analyzing combinations of mutations that span from ancestral to derived sequences and indicate that molecular mechanisms in evolution can vary depending on the context. Further studies of different molecules in different contexts will likely provide important insights into commonalities in molecular evolution as well as structural and biochemical features that mediate distinctions for different molecules. Systematic mutational analyses promise to contribute greatly to this area as it is currently feasible with sequencing-based readouts to analyze the function of ~100,000 separate sequences (Melamed *et al.* 2013), which is theoretically sufficient to monitor all possible combinations between two possible mutations at 17 positions ($2^{17} \approx 10^5$).

Directed-evolution approaches have provided insights into the landscape of mutations near functional peaks of fitness. Directed evolution (Oliphant and Struhl 1989) can identify highly functional variants from a library of gene variants and it has provided important insights into the landscape of available adaptive mutations (Figure 1C). The contributions of directed evolution toward understanding adaptive evolution have recently been reviewed (Bloom and Arnold 2009) and include the observation that mutations that increase folding stability can play an important role in adaptation by increasing the tolerance to secondary adaptive mutations that are destabilizing. Directed evolution has also demonstrated that mutations that increase biochemical promiscuity (decreasing specificity) can, under further selection, lead to new specialized function. Both directed evolution and ancestral reconstruction approaches have revealed an important role for permissive mutations in the evolution of new function. These permissive mutations maintain the ancestral function while at the same time enabling subsequent mutations to have a stronger impact on function. These studies demonstrate that the effects of mutations can depend strongly on secondary mutations within the same gene.

The effects of mutations can also depend strongly on secondary mutations in other genes. Genes that mediate protein homeostasis, especially chaperones, can strongly influence the effects of mutations in many other genes. In eukaryotes, the Hsp90 chaperone has been shown to broadly influence the impact of mutations in other genes (Jarosz *et al.* 2010). Normal levels of Hsp90 provide excess chaperone capacity that masks the effects of many mutations to Hsp90 clients that would experience folding defects were Hsp90 function reduced. Consistent with this idea, reducing Hsp90 levels reveals phenotypic effects for many otherwise “cryptic” mutations in yeast (Cowen and Lindquist 2005). In bacteria, the level of the chaperonin GroEL/ES plays a similar role and masks the effects of many mutations in client proteins (Tokuriki and Tawfik 2009a). In their study, overexpression of GroEL/ES provided a roughly twofold increase in the probability that random mutations in three different client proteins would retain function. These studies and many others clearly demonstrate that the interdependence of mutations in different genes can have a large influence on evolution.

Figure 1 Conceptual depictions and interpretations of mutational landscapes. (A) Rendition of a landscape after those envisioned by Wright (1932) to conceptualize vast possible combinations of different alleles. Allele space is depicted on the image plane and contour lines indicate relative fitness. (B) Modern molecular analyses of mutational steps between ancestral and derived sequences have provided insights into available pathways on mutational land-

scapes. (C) Peaks of fitness or function under defined conditions have been explored using approaches including directed evolution, that can identify a small set of highly functional variants from stochastically generated mutations. (D) Systematic maps of local regions of mutational space (e.g., all single-step mutations from the parental sequence of a gene) can be generated using current sequencing-based approaches. These depictions are intended to provide a conceptual outline, but they are inaccurate in detail (e.g., the vastness of possible mutational space is not accurately represented, and the smoothness of the surfaces does not accurately represent the observation that single-step mutations can lead to dramatic cliff-like changes in fitness).

Quantification of Local Mutational Landscapes Using Sequencing-Based Approaches

Modern sequencing approaches can quantify the response of thousands of mutations to selective pressures (Fowler *et al.* 2010; Hietpas *et al.* 2011), providing detailed views into local regions of mutational space (Figure 1D). These approaches take advantage of the power of bulk competitions to measure the relative biochemical effects (e.g., binding affinity for a partner protein) or physiological effects (e.g., rate of cell proliferation) of many different mutations in a single experiment (Figure 2A). The capability to sequence millions of individual DNA molecules from a complex sample provides the opportunity to accurately measure the frequency of thousands of mutations both before and after a selection event (Figure 2A). The change in frequency of a mutation due to selection is a direct measure of the functional effects of that mutation compared to other variants in the bulk competition. The dynamic range of functional effects that can be monitored in bulk competitions depends on the experimental setup. The strength of highly deleterious mutations can be effectively distinguished with a minimum level of effective selection (e.g., short times in growth competition). In contrast, mutations with small effects relative to the wild type (WT) distinguish themselves only with high levels of effective selection (e.g., after many generations in growth competition). By analyzing the frequency of mutations at multiple levels of effective selection, bulk competitions can distinguish mutations of both small and large experimental effect.

Bulk competitions where many hundreds or thousands of mutations are analyzed in the same physical sample ensure that each mutation experiences equivalent experimental conditions, which facilitate precise measurements of the effects of each mutation (Hietpas *et al.* 2013b). Mutations in genes that impact cooperation among individuals often lead to distinct physiological impacts in monoculture vs. bulk cultures. For example, *Saccharomyces cerevisiae* that do not produce the enzyme that hydrolyzes sucrose cannot grow on this sugar source in monoculture, but can exhibit a fitness advantage

when grown in a co-culture with yeast that do produce this enzyme (Gore *et al.* 2009). This advantage occurs because sucrose hydrolysis occurs outside the cell membrane such that the products of sucrose hydrolysis are available to the entire culture, including the cheater yeast that benefit from sucrose without having to invest in producing the hydrolase enzyme. However, for most genes that have been subjected to systematic mutational analyses, the effects of mutations in bulk cultures have correlated strongly with effects of individual mutations analyzed in isolation (Hietpas *et al.* 2011; Roscoe *et al.* 2014). In addition to environmental conditions, experimental reproducibility of bulk competitions depends on many factors, including counting robustness (e.g., due to sequencing depth) and population management (e.g., bottlenecks where population size approaches mutational diversity will lead to stochastic frequency changes from random sampling). With careful attention to these issues (Hietpas *et al.* 2013a), full experimental repeats of bulk competitions monitored by sequencing exhibit strong reproducibility and are capable of distinguishing functional effects on the order of 0.1% (Hietpas *et al.* 2013b; Bank *et al.* 2014).

There are many potential approaches to generate systematic libraries of mutations, and each approach has distinct implications for the type of conceptual issues that are best addressed. Pioneering work by Fowler *et al.* (2010) utilized gene synthesis with engineered degeneracy at many consecutive nucleotide positions to generate libraries containing the majority of single-nucleotide substitutions, a fraction of possible double nucleotide substitutions, along with smaller fractions of higher-order nucleotide substitutions. This approach, coined deep mutational scanning, provides outstanding opportunities to investigate the interdependency between mutations in the same gene (Araya *et al.* 2012). A distinct approach (Hietpas *et al.* 2011) termed “exceedingly methodical and parallel investigation of randomized individual codons” (EMPIRIC), was developed to analyze all possible single-amino-acid substitutions (including one, two, or three nucleotide substitutions at each codon). Similar approaches to analyze all possible amino

Figure 2 Sequencing-based approach to quantify the frequency of mutations in bulk competitions. (A) Engineered mutations in bulk libraries are subjected to selection pressure. Sequencing of samples before and after selection provides a direct estimate of the frequency change of each mutation, which is a measure of function or fitness under the conditions of the experiment. (B) Contributions of read depth to estimates of mutational frequencies. The fraction of a mutation (F_i) is defined by the number of copies of that mutation (N_i) relative to total variants in the sample (N_{total}). In sequencing approaches, the number of reads of a mutation (R_i) relative to total reads (R_{total}) provides an estimate of the frequency of mutations (E_i). These estimated mutational frequencies contain experimental noise. Many factors contribute to experimental noise, including sequencing errors and read depth. Read depth is frequently a dominant source of noise. Based solely on sampling, the standard error (S_p) describes the expected noise in estimating the frequency of a mutation based on read depth. The probability of an estimate falling within two standard errors from the true value is 95%. The graph illustrates how read depth (x-axis) impacts the noise in estimating mutational frequencies (y-axis). This graph is for a mutation at 0.1% frequency (for mutations at frequencies $<1\%$, the relationship is similar). (C) Measuring multiple points during selection (e.g., at multiple time points in a growth competition), can reduce the impact of read noise for any individual point on fitness estimates. This graph shows observations of the effects of mutations in Hsp90 (N588P^{CCC} in green, a silent N588A^{ACC} mutation in gray, and a nonsense N588*^{TGA} mutation in red) on yeast growth in an elevated saline environment (Hietpas *et al.* 2013b). (D) The frequency of mutations can be estimated by directly sequencing regions containing mutations. (E) Alternatively, a barcode can be introduced outside of the ORF and associated with mutations in the ORF using tiled paired-end reads. The barcode can then be efficiently sequenced to infer the frequency of the associated ORF mutations. Barcode strategies require additional setup, but have two important advantages: they enable analyses of mutations that are separated by large distances (e.g., greater than can be spanned in a single sequencing reaction), and they enable error correction for indexes that differ by more than one base from all others.

acid substitutions have been independently developed (Fleishman *et al.* 2011; McLaughlin *et al.* 2012; Whitehead *et al.* 2012). These EMPIRIC-style approaches are well suited to investigate biophysical requirements that can be revealed by surveying the effects of all 20 natural amino acids at each position in a protein. By focusing on single-amino-acid substitutions, EMPIRIC-style libraries are efficient to analyze and provide promising opportunities to investigate additional dimensions, including organism-level genetic background (e.g., mutations in other genes) and varied environmental conditions.

Two additional experimental factors that greatly influence the scale and precision of sequencing readouts from bulk competitions warrant discussion: read depth and read errors. The number of times that a mutation is read (observed in sequencing) impacts noise from sampling. Excluding other factors, the noise from this type of sampling is defined by the standard error function (Figure 2B) and indicates that counting a mutation 400 times will yield a frequency estimate within 10% of the true value ($P > 0.95$ from binomial expectations). The relationship between read depth and sampling noise is nonlinear (Figure 2B), leading to rapidly diminishing returns from further increases in sequencing coverage. More precise estimates of the effects of mutations on selection can be determined by analyzing the frequency change of mutations at multiple levels of selection. For example, monitoring

the frequency of mutations at multiple time points in a bulk growth competition can provide a more precise estimation of fitness effects than estimates from just two time points (Figure 2C).

Read errors can distort analyses of mutation frequencies and multiple strategies have been utilized to address this issue. Early strategies (Figure 2D) included the analysis of regions of DNA short enough such that both strands could be sequenced in the same molecule, resulting in double interrogation of each base pair (Fowler *et al.* 2010). The implementation of efficient strategies to generate libraries of predominantly point mutations (Hietpas *et al.* 2011) enable most read errors to be identified as apparent double-amino-acid substitutions and filtered from further analyses. A recent study has taken advantage of a powerful indexing strategy (Starita *et al.* 2013) where stretches of ~ 20 bases of random sequence “barcodes” were added outside of the open reading frame (ORF) and associated with mutations in the ORF (Figure 2E). These barcodes were associated with mutations in the open reading frame using paired-end reads with the barcode at one end and tiled regions of the ORF at the other end (Hiatt *et al.* 2010). Once associated, the relative abundance of a mutation in the open reading frame could be determined by measuring the frequency of barcodes. In this setup, the potential complexity of the barcode (e.g., $4^{20} \approx 10^{12}$ for a 20-base random

region) is such that each barcode is distinct at multiple bases from every other barcode sequence in the library. This clever approach is extremely powerful because the barcodes can be proofread (e.g., most misreads in the barcodes can be detected and corrected) and because the frequency of mutations spread across large open reading frames can be determined by sequencing a short barcode region. Indexing approaches are appealing for systems where indices can be readily incorporated with libraries of mutations and where mutational sampling is not limiting.

Structural Interpretations of Local Mutational Landscapes

Systematic analyses of the effects of individual amino acid substitutions on function have revealed local mutational landscapes, which are often visualized as heatmaps (Figure 3A) (Fowler *et al.* 2010; Hietpas *et al.* 2011; McLaughlin *et al.* 2012; Whitehead *et al.* 2012; Melamed *et al.* 2013; Starita *et al.* 2013; Hsu *et al.* 2014; Roscoe *et al.* 2014). Mapping functional effects onto structures of these proteins has highlighted some features that appear common. For example, most amino acid positions on the surface of proteins have been found either very tolerant (where most substitutions have no observable impact on function) or very sensitive (where most substitutions dramatically reduce protein function). In most of these systems where binding sites have been structurally characterized, the sensitive positions cluster at direct interfaces (Figure 3, B and C). These observations are consistent with long-standing inferences that contact surfaces on proteins should impose strong evolutionary constraints (King and Jukes 1969; Zuckerkandl 1976). In contrast, hydrophobic amino acids located in the solvent inaccessible core of proteins commonly tolerate substitutions that maintain hydrophobic characteristics (Rennell *et al.* 1991; McLaughlin *et al.* 2012; Roscoe *et al.* 2014). These observations are consistent with the strength and relatively nonspecific characteristic of hydrophobic interactions (Dill 1990) and previous analyses of smaller sets of variants (Cordes *et al.* 1996). From a structural perspective, the novel aspect of sequencing-based approaches is in the comprehensiveness of the analyses. Comprehensive maps of mutations provide new opportunities to investigate connections between physics, biochemistry, and physiology.

In almost every experimentally studied protein, the effects of point mutations on function are predominantly bimodal, with most mutations causing either undetectable changes to function or severe defects. This has been observed in systematic analyses of local mutational landscapes (Wylie and Shakhnovich 2011; Jiang *et al.* 2013), analyses of sparsely sampled random mutations (Sanjuan *et al.* 2004), and a variety of clever approaches to interrogate systematically engineered mutations developed prior to the availability of high-throughput sequencing readouts (Cunningham and Wells 1989; Rennell *et al.* 1991; Palzkill and Botstein 1992). Bimodal distributions of fitness effects are common to many different genes and systems (Rennell *et al.* 1991; Sanjuan

et al. 2004; Domingo-Calap *et al.* 2009; Jiang *et al.* 2013). Bimodal fitness effects highlight the rugged or discontinuous properties of fitness landscapes as dramatic fitness changes are commonly observed for the smallest mutational step. A single substitution can, and often does, push protein function off a cliff, which makes mutational landscapes rough in character (Figure 3A). The prevalence of strongly deleterious mutations is one of the main determinants of selection for high fidelity in genome replication, including low error rates for most replicative polymerases (Lynch 2010). The near universality of bimodal fitness effects of mutations and their strong influence on evolution has motivated many efforts to understand their biochemical and biophysical underpinnings. A handful of different mechanisms (Tokuriki and Tawfik 2009c; Bershtein *et al.* 2013; Jiang *et al.* 2013) have been shown to contribute to bimodal fitness effects and the relative influence of each mechanism can vary dramatically depending on the protein or context. Each mechanism that leads to bimodal fitness effects has a distinct nonlinear relationship at its core. These include nonlinear relationships between thermodynamic folding stability and the ratio of folded to unfolded protein (Tokuriki and Tawfik 2009c), complex relationships between protein sequence and degradation rates in cells (Bershtein *et al.* 2013), and nonlinear relationships between biochemical (e.g., enzyme proficiency) and physiological function (Kacser and Burns 1981; Lunzer *et al.* 2005; Jiang *et al.* 2013; Roscoe and Bolon 2014),

The influence of cooperative protein folding on evolution should apply to all proteins whose function requires a well-defined native structure (Tokuriki and Tawfik 2009c). Based on thermodynamic arguments, selection will favor proteins that predominantly populate folded states. However, once a protein is sufficiently stable to predominantly populate the native state, further increases in stability will have only marginal impacts on function. Because mutations that destabilize proteins are more common than stabilizing mutations (Baase *et al.* 1997), mutation and drift tend to hinder the evolution of hyperstable proteins in the absence of other considerations. In addition to drift, hyperstable proteins can also be selected against, if excess stability interferes with flexibility important for function or the ability to evolve new functions (Tokuriki and Tawfik 2009b). As with many questions regarding evolutionary mechanism, it is difficult to distinguish the relative contributions of drift and selection with regards to protein folding stability (Lynch *et al.* 2011).

The influence of protein degradation rates on fitness depends on association with cellular proteases, a property that can be distinct from protein folding stability. The proteases that appear responsible for most protein degradation in cells are compartmentalized such that the active sites are sequestered in a large internal cavity (Baker and Sauer 2012). Gatekeeper proteins bind to specific recognition determinants in substrates and hydrolyze ATP to drive the transport of bound substrates into the degradation chamber. The rate of substrate degradation is determined in large part by binding affinity for the gatekeeper complex, which may

Figure 3 Local mutational landscapes mapped to structure. (A) Heat map representation of a functional landscape for ubiquitin where all individual amino acid substitutions were analyzed (Roscoe and Bolon 2014). For each amino acid position, the average impact of substitutions (bottom row) is a representation of mutational sensitivity. (B) Mapping the effects of ubiquitin mutations to structure (Peschard *et al.* 2007) indicates that surfaces that directly contact binding partners (the purple ribbon illustrates a binding partner) are sensitive, while distal surfaces are very tolerant of amino acid substitutions. (C) Similar patterns of mutational sensitivity at binding interfaces have also been observed in systematic analyses of an RRM domain (Melamed *et al.* 2013) and a WW domain (Fowler *et al.* 2010). In B and C, sensitive positions where the average effect of a mutation was strongly deleterious are colored blue, while tolerant positions are colored yellow, intermediate positions, light blue, and binding partners, purple.

not directly relate to the folding stability of substrates. In Arc repressor, the selection for suppressors of a mutation that impaired the folding stability identified sequence changes in the unstructured C terminus that did not impact folding stability, but instead reduced interactions with cellular proteases (Bowie and Sauer 1989). A recent study of the fitness effects of mutations in dihydrofolate reductase (DHFR) in *Escherichia coli* found that the mutant phenotype caused by most deleterious mutations could be suppressed by deletion of the gene encoding the Lon protease (Bershtein *et al.* 2013). Because Lon should not impact the folding stability of DHFR, their study demonstrates that protein quality control processes that are not directly correlated with folding stability can have a strong influence on the fitness effects of mutations. Their study indicates that many factors influence protein degradation rates in cells and that protein stability to unfolding can be a poor predictor of degradation susceptibility.

While less studied than protein stability, biochemical flux relationships also appear common and have the potential to strongly influence the effects of mutations. The relationships between fluxes in biochemical pathways (e.g., the rate of product formation for an enzyme) and physiological function (e.g., growth rate) are commonly nonlinear such that the expression level of many essential proteins can be reduced dramatically without compromising fitness. Recent analyses of the essential Hsp90 chaperone in yeast demonstrated that the expression level of this protein must be reduced ~100-fold to reduce growth rate 2-fold under standard conditions (Jiang *et al.* 2013). Studies of Hsp90 at multiple expression levels demonstrated that many mutations that caused strong fitness defects at limiting expression did not cause observable defects at endogenous expression levels (Jiang *et al.* 2013).

Evolvability is a consideration that has also been examined with regard to thermodynamic protein stability. Proteins with higher thermodynamic stability can tolerate more mutations without unfolding. For this reason, proteins with strong thermodynamic stability have been found to promote adaptation in directed evolution experiments (Bloom *et al.* 2006). Based on this logic, mutations that increase thermodynamic stability can be broadly permissive by increasing the tolerance of the protein to many secondary mutations that would result in unfolding in less thermodynamically stable sequence backgrounds (Gong *et al.* 2013). Mutations that increased thermodynamic stability were recently identified from a systematic analysis of a large set of multiple mutations in a WW domain based on their capability to rescue the peptide binding function of secondary mutations that had binding defects in other sequence backgrounds (Araya *et al.* 2012).

In addition to global conformational transitions between native and unfolded conformations, many studies have also highlighted critical functional contributions from subtle changes in protein conformation and dynamics. Some of the most fundamental observations of protein conformational dynamics and their contributions to function come from NMR analyses of enzymes where highly specific protein motions can be required for efficient catalysis (Eisenmesser *et al.* 2002; Villali and Kern 2010). Recent systematic analyses of mutations in proteins or regions of proteins involved in protein-protein interactions (Jiang *et al.* 2013; Lee *et al.* 2014; Roscoe and Bolon 2014; Roscoe *et al.* 2014) indicate that binding functions can also be very sensitive to protein conformational changes less severe than complete unfolding. This inference comes largely from observations that many amino acid substitutions at solvent inaccessible positions in these

binding proteins cause strong functional defects without causing protein unfolding. A striking example of this comes from studies of ubiquitin, which by many measures is one of the least dynamic proteins. Ubiquitin is a rock in terms of temperature stability and at neutral pH requires temperatures in excess of boiling to unfold (Wintrode *et al.* 1994). In addition, ubiquitin primarily acts through protein–protein binding interactions. Despite these features, many mutations in the core of ubiquitin that cause subtle alterations to the conformational properties of ubiquitin have large impacts on its binding functions (Phillips *et al.* 2013; Lee *et al.* 2014; Roscoe *et al.* 2014). Further studies on additional proteins should provide valuable insights into the generality of these observations for binding proteins and the nature of biophysical constraints in these systems.

Evolutionary Inferences From Comparisons of Experimental Fitness Effects to Sequence Divergence Observed in Nature

For many functionally conserved proteins, the conservation of amino acids observed in nature is stronger than would be predicted based on simple interpretations of the effects of mutations in laboratory experiments (Hietpas *et al.* 2011; Melamed *et al.* 2013; Starita *et al.* 2013; Roscoe *et al.* 2014). These studies find that many mutations that were compatible with full experimental function (within the limits of detection) were not observed in alignments of sequences from natural isolates (Figure 4A). Among the possible explanations for these observations, two main themes stand out: limited environmental conditions are explored in laboratory experiments compared to nature, and mutations with fitness effects that are too small to measure in laboratory experiments can be subject to natural selection in large populations over evolutionary time scales. These explanations are not mutually exclusive and both likely contribute to distinctions between selection in laboratory experiments and nature.

The ability to control environmental conditions in laboratory selections is valuable for exploring how distinct conditions impact the effects of mutations (McLaughlin *et al.* 2012; Hietpas *et al.* 2013b; Bank *et al.* 2014), but sampling all possible conditions that could be experienced in nature is impractical. For this reason, there is great promise in combining ecological studies that identify relevant conditions in nature with laboratory investigations that sample how these conditions impact mutational landscapes.

The effective population size (N_e) determines the influence of stochastic processes, including drift, on the frequency of a mutation in nature. As effective population size decreases, the influence of drift increases. For a mutation to escape drift, the fitness change of the allele relative to others must be of sufficient magnitude such that selection has a stronger influence than drift. Population-genetic theory indicates that the effects of mutations subject to selection in natural populations are proportional to the inverse of effective population size (Ohta 1973). The near-neutral window describes mutations

with sufficiently small fitness effects that their frequency is primarily mediated by drift. If the effect of a deleterious mutation is larger in magnitude than the near-neutral cutoff ($\approx 1/N_e$) it is likely to be purged from the population. Similarly, if the effect of a beneficial mutation is stronger than this cutoff, it is likely to consistently increase in frequency in the population. The effective population size, which is the number of breeding individuals in a population, determines the distribution of allele frequencies and can be estimated from sequencing many individuals in a population (Lynch and Conery 2003). The effective population size of most microbes is very large, with estimates for the yeast *S. paradoxus* on the order of 10^6 (Tsai *et al.* 2008). Fitness effects on the order of $\sim 0.0001\%$ for yeast with population demographics similar to *S. paradoxus* have a high probability to escape drift and be subject to natural selection.

For a protein whose function is conserved in nature, the predominant form of selection would likely be to maintain function. In this case, mutations with very small deleterious effects (e.g., 0.1%) and those with strong deleterious effects (e.g., null) would both be efficiently purged from large natural populations over long time scales. For this reason, the strength of deleterious mutations is difficult to infer from analyses of natural sequences (Figure 4B). In contrast, experimental approaches can delineate fitness effects across broad spans (from null to beneficial), but cannot readily distinguish fitness effects to the level of resolution necessary to identify mutations that would fall within the near-neutral ($\approx 1/N_e$) window for most natural populations (Figure 4B). Experimental approaches have been developed to measure small changes in protein function by increasing the stringency of selection (Jiang *et al.* 2013; Roscoe and Bolon 2014). The engineering required to increase selection stringency may impose artificial distortions of large magnitude relative to the width of the near-neutral window. Inferences regarding the near-neutral window based on artificially increased selection stringency should be made with caution.

This rationale has important implications for interpreting conservation patterns in alignments of related proteins. Conservation at a position in an alignment is an indication of purifying selection. However, the strength of purifying selection at conserved positions is unclear because mutations that have deleterious effects greater in magnitude than the near-neutral window are purged from populations in a similar fashion. For example, a mutation that causes a 1% fitness defect and a null mutation with a 100% fitness defect will both be efficiently purged from large natural populations. For this reason, it is difficult to infer the strength of deleterious mutations from alignments of related proteins. The strength of deleterious mutations can be directly measured in laboratory investigations. Comparisons of sequence conservation in nature and systematic experimental analyses provide information that is highly complementary. For example, sequence conservation in ubiquitin is strong across the eukaryotic lineage with only four amino acid positions exhibiting variation, indicating that the entire protein is

Figure 4 Comparison between conservation patterns in nature and experimental fitness measurements. (A) Heat map representation of both amino acids observed for a region of the Hsp90 protein from diverse eukaryotes and experimental fitness effects observed for the same region of yeast Hsp90. The function of Hsp90 is strongly conserved in eukaryotes (Picard *et al.* 1990), suggesting that selection acting on Hsp90 may be predominantly purifying in nature. (B) Illustration of inferences that can be made from sequence divergence patterns for genes subject to predominantly purifying selection compared to experimental measurements of fitness effects. Slightly deleterious (e.g., 1% defects) and strongly deleterious (e.g., null) mutations should both be efficiently purged from large natural populations over evolutionary time scales.

The breadth of the near-neutral window of fitness effects in nature is theoretically proportional to the inverse of effective population size (Ohta 1973). Experimental approaches can monitor the effects of mutations across the full breadth of fitness (null to beneficial), but the resolution is not sufficient to distinguish the window that would be near neutral in large natural populations over long time scales. Studies of experimental fitness and patterns of divergence in nature provide distinct information that, analyzed together, can be more powerful than either approach alone.

subject to purifying selection. However, the magnitude of selection at each position is unclear from these patterns of conservation. Systematic analyses of ubiquitin mutations in laboratory experiments demonstrate a great variation in the fitness effects at conserved positions (Roscoe *et al.* 2014). While structurally characterized binding interfaces on ubiquitin were very sensitive to mutation (most substitutions caused null fitness), all other surface positions were very tolerant (most mutations do not cause a distinguishable impact on fitness). The experimentally tolerant yet naturally conserved positions have two potential explanations: fitness effects subject to selection in large natural populations are beyond the resolution power of current experiments, and/or diverse environmental conditions experienced in nature may lead to more complex and stringent selection. These observations highlight the value of systematic mutation approaches to distinguish biochemical hotspots that cannot be easily discerned from analyses of natural sequences alone.

The potential interdependence of mutations is another important factor to consider in comparing laboratory investigations of mutations to conservation/divergence patterns in nature. Mutations that have interdependent effects are often termed epistatic. In contrast, nonepistatic or independent mutations have similar fitness effects in different genetic backgrounds. Understanding the relative contributions of independent and epistatic mutations to evolution is an active area of current theoretical and experimental studies (Breen *et al.* 2012; Gong and Bloom 2014). Mutations whose effects are largely independent or epistatic should lead to distinct patterns in comparisons between laboratory experiments and divergence patterns in nature. Because selection in natural systems is generally stringent (e.g., due to large effective population sizes), the observation of a mutation at high frequency in a natural lineage is an indication that the mutation is highly fit in the genetic background where it was observed.

If the effect of a naturally observed mutation is largely independent, it follows that this mutation will likely be functional in the genetic background utilized in laboratory experiments. While further studies along this line are warranted, initial studies indicate that most mutations observed in nature exhibit high function in genetic backgrounds that have been examined in laboratory experiments (Hietpas *et al.* 2011; Melamed *et al.* 2013; Roscoe *et al.* 2014). In contrast, if the natural mutation is strongly epistatic, then it may exhibit a defect in the laboratory experiments. By this logic, comparisons of mutational landscapes from laboratory experiments to alignments of sequences from nature provide potentially powerful approaches to examine the relevant contribution of independent and epistatic mutations to the evolution of different genes. Recent studies have noted that the relative role of epistasis in natural evolution is highly variable depending on the gene and selective pressures in nature (Gong and Bloom 2014). Systematic mutation experiments promise to add to this active area of study.

Probing How Complex and Interconnected Biological Systems Contribute to Cell Physiology

Complex networks of macromolecular interactions (e.g., phosphorylation cascades) mediate many physiological processes. The organization of proteins in these networks can have a strong influence on the interdependence of double knockouts. For example, the knockout of any protein in a phosphorylation cascade should block the signaling pathway, such that the knockout of additional proteins in the cascade would not further alter signaling. Motivation to understand interaction networks has contributed to the development of powerful biochemical (Walzthoeni *et al.* 2013) and genetic approaches (Chien *et al.* 1991) to map macromolecular interactions. Indeed, systems biology has emerged as a field largely to

understand how interaction networks underlie cell physiology (Sahni *et al.* 2013). In this field, systems of genes or gene products are often considered as nodes with interactions (e.g., direct protein–protein binding) between these nodes described as edges (Figure 5).

Systematic analyses of the effects of mutations on specific biochemical function and measures of cell physiology (e.g., growth rate) provide promising avenues to enhance our understanding of the biochemical networks that underlie complex biological systems (Figure 5B). Many studies have shown that biochemical screens can accurately quantify how thousands of mutations impact the strength of biochemical interactions, including protein–protein binding (Fowler *et al.* 2010; McLaughlin *et al.* 2012; Roscoe and Bolon 2014). Analogous approaches can also quantify the effects of the same set of thousands of mutations on physiological functions, including growth rate (Roscoe and Bolon 2014). Combined analyses of the effects of mutations on biochemical properties and physiological function provide the opportunity to train quantitative flux models of cell physiology (Powers *et al.* 2012).

Initial studies of this type have been undertaken for the ubiquitin system. The impacts of all ubiquitin point mutations were analyzed for biochemical activation by the E1 protein (Roscoe and Bolon 2014) and compared to impacts of the same mutations on yeast growth rate. Ubiquitin binds to hundreds of different partner proteins in cells. Even so, this study could distinguish ubiquitin mutations that primarily impacted E1 activation. These inferences were based on the logic that among mutations with similar E1 defects, those that primarily impacted E1 would have the least severe physiological defects. The results of this study indicated that E1 activation efficiency could be reduced dramatically (~50-fold) without compromising yeast growth. This study highlights the potential of systematic mutational analyses to provide insights into the connections between biochemical properties and physiological outcomes.

Concluding Thoughts

Systematic analyses of mutations provide a technological advancement with the potential to improve the resolution of experiments across many biological disciplines. Sequencing readouts enable the precise quantification of biochemical and physiological effects of thousands of mutations in parallel. This is one of a growing number of approaches that enable and promote cross-disciplinary biological research (Harms and Thornton 2013). Pioneering approaches to sequence virus populations with extreme depth have recently provided valuable views of the landscape of mutation effects across a genome (Acevedo *et al.* 2014). This genome-wide study of poliovirus showed a bimodal distribution of fitness effects with a large cluster of mutations that were of minimal impact and a second cluster that were null. Intriguingly, 10% of silent mutations were observed to be lethal. This study also quantified mutational rates and demonstrated that for different mutations (e.g., G→A or A→C) these rates spanned more

Figure 5 Relationships between biochemical and physiological function. This illustration shows a map of a protein–protein interaction network where the colored circles or nodes represent distinct proteins and the lines or edges connecting them represent binding interactions. Systematic analyses of the effects of mutations on both specific biochemical functions (e.g., strength of binding between two proteins) and overall contributions to physiological function provide new opportunities to interrogate complex biological networks that mediate many critical processes.

than two orders of magnitude. Systematic analyses of engineered mutations have been extended beyond microbes to investigate drug resistance in mammalian cells (Wagenaar *et al.* 2014). This study of the V600E BRAF oncogene systematically identified mutations in the kinase active site that enable cultured cells to grow in the presence of drug therapy. Given the continuing technological developments in sequencing-based approaches, it appears likely that highly precise measurements of the effects of mutations across entire genomes (at least for viruses) in laboratory experiments are on the immediate horizon. These technical advancements provide great opportunities for enhancing our understanding throughout many different biological disciplines, including biochemistry, cell physiology, and evolution.

Literature Cited

- Acevedo, A., L. Brodsky, and R. Andino, 2014 Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 505: 686–690.
- Araya, C. L., D. M. Fowler, W. Chen, I. Muniez, J. W. Kelly *et al.*, 2012 A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA* 109: 16858–16863.
- Baase, W. A., L. Liu, D. E. Tronrud, and B. W. Matthews, 1997 Lessons from the lysozyme of phage T4. *Protein Sci.* 19: 631–641.
- Baker, T. A., and R. T. Sauer, 2012 ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta* 1823: 15–28.
- Bank, C., R. T. Hietpas, A. Wong, D. N. Bolon, and J. D. Jensen, 2014 A bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. *Genetics* 196: 841–852.

- Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59.
- Bershtein, S., W. Mu, A. W. Serohijos, J. Zhou, and E. I. Shakhnovich, 2013 Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol. Cell* 49: 133–144.
- Bloom, J. D., and F. H. Arnold, 2009 In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. USA* 106(Suppl 1): 9995–10000.
- Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold, 2006 Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA* 103: 5869–5874.
- Bowie, J. U., and R. T. Sauer, 1989 Identification of C-terminal extensions that protect proteins from intracellular proteolysis. *J. Biol. Chem.* 264: 7596–7602.
- Breen, M. S., C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, 2012 Epistasis as the primary factor in molecular evolution. *Nature* 490: 535–538.
- Chien, C. T., P. L. Bartel, R. Sternglanz, and S. Fields, 1991 The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* 88: 9578–9582.
- Cordes, M. H., A. R. Davidson, and R. T. Sauer, 1996 Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* 6: 3–10.
- Cowen, L. E., and S. Lindquist, 2005 Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. *Science* 309: 2185–2189.
- Cunningham, B. C., and J. A. Wells, 1989 High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* 244: 1081–1085.
- Dill, K. A., 1990 Dominant forces in protein folding. *Biochemistry* 29: 7133–7155.
- Domingo-Calap, P., J. M. Cuevas, and R. Sanjuan, 2009 The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. *PLoS Genet.* 5: e1000742.
- Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle *et al.*, 2009 Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133–138.
- Eisenmesser, E. Z., D. A. Bosco, M. Akke, and D. Kern, 2002 Enzyme dynamics during catalysis. *Science* 295: 1520–1523.
- Fleishman, S. J., T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn *et al.*, 2011 Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332: 816–821.
- Fowler, D. M., C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany *et al.*, 2010 High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7: 741–746.
- Gong, L. I., and J. D. Bloom, 2014 Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS Genet.* 10: e1004328.
- Gong, L. I., M. A. Suchard, and J. D. Bloom, 2013 Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2: e00631.
- Gore, J., H. Youk, and A. van Oudenaarden, 2009 Snowdrift game dynamics and facultative cheating in yeast. *Nature* 459: 253–256.
- Harms, M. J., and J. W. Thornton, 2013 Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* 14: 559–571.
- Hiatt, J. B., R. P. Patwardhan, E. H. Turner, C. Lee, and J. Shendure, 2010 Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7: 119–122.
- Hietpas, R., B. Roscoe, L. Jiang, and D. N. Bolon, 2013a Fitness analyses of all possible point mutations for regions of genes in yeast. *Nat. Protoc.* 7: 1382–1396.
- Hietpas, R. T., J. D. Jensen, and D. N. Bolon, 2011 Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. USA* 108: 7896–7901.
- Hietpas, R. T., C. Bank, J. D. Jensen, and D. N. Bolon, 2013b Shifting fitness landscapes in response to altered environments. *Evolution* 67: 3512–3522.
- Hsu, H. J., K. H. Lee, J. W. Jian, H. J. Chang, C. M. Yu *et al.*, 2014 Antibody variable domain interface and framework sequence requirements for stability and function by high-throughput experiments. *Structure* 22: 22–34.
- Ingolia, N. T., S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, 2009 Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218–223.
- Jarosz, D. F., M. Taipale, and S. Lindquist, 2010 Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. *Annu. Rev. Genet.* 44: 189–216.
- Jiang, L., P. Mishra, R. T. Hietpas, K. B. Zeldovich, and D. N. Bolon, 2013 Latent effects of Hsp90 mutants revealed at reduced expression levels. *PLoS Genet.* 9: e1003600.
- Kacser, H., and J. A. Burns, 1981 The molecular basis of dominance. *Genetics* 97: 639–666.
- King, J. L., and T. H. Jukes, 1969 Non-Darwinian evolution. *Science* 164: 788–798.
- Lee, S. Y., L. Pullen, D. J. Virgil, C. A. Castaneda, D. Abeykoon *et al.*, 2014 Alanine scan of core positions in ubiquitin reveals links between dynamics, stability, and function. *J. Mol. Biol.* 426: 1377–1389.
- Lunzer, M., S. P. Miller, R. Felsheim, and A. M. Dean, 2005 The biochemical architecture of an ancient adaptive landscape. *Science* 310: 499–501.
- Lynch, M., 2010 Evolution of the mutation rate. *Trends Genet.* 26: 345–352.
- Lynch, M., and J. S. Conery, 2003 The origins of genome complexity. *Science* 302: 1401–1404.
- Lynch, M., L. M. Bobay, F. Catania, J. F. Gout, and M. Rho, 2011 The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet.* 12: 347–366.
- Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.
- McLaughlin, Jr., R. N., F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, 2012 The spatial architecture of protein function and adaptation. *Nature* 491: 138–142.
- Melamed, D., D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, 2013 Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* 19: 1537–1551.
- Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha *et al.*, 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344–1349.
- Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96–98.
- Oliphant, A. R., and K. Struhl, 1989 An efficient method for generating proteins with altered enzymatic properties: application to beta-lactamase. *Proc. Natl. Acad. Sci. USA* 86: 9094–9098.
- Ortlund, E. A., J. T. Bridgman, M. R. Redinbo, and J. W. Thornton, 2007 Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317: 1544–1548.
- Palzkill, T., and D. Botstein, 1992 Probing beta-lactamase structure and function using random replacement mutagenesis. *Proteins* 14: 29–44.
- Peschard, P., G. Kozlov, T. Lin, I. A. Mirza, A. M. Berghuis *et al.*, 2007 Structural basis for ubiquitin-mediated dimerization and

- activation of the ubiquitin protein ligase Cbl-b. *Mol. Cell* 27: 474–485.
- Phillips, A. H., Y. Zhang, C. N. Cunningham, L. Zhou, W. F. Forrest *et al.*, 2013 Conformational dynamics control ubiquitin-deubiquitinase interactions and influence in vivo signaling. *Proc. Natl. Acad. Sci. USA* 110: 11379–11384.
- Picard, D., B. Khursheed, M. J. Garabedian, M. G. Fortin, S. Lindquist *et al.*, 1990 Reduced levels of hsp90 compromise steroid receptor action in vivo. *Nature* 348: 166–168.
- Powers, E. T., D. L. Powers, and L. M. Gierasch, 2012 FoldEco: a model for proteostasis in *E. coli*. *Cell Reports* 1: 265–276.
- Rennell, D., S. E. Bouvier, L. W. Hardy, and A. R. Poteete, 1991 Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* 222: 67–88.
- Roscoe, B. P., and D. N. Bolon, 2014 Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J. Mol. Biol.* 426: 2854–2870.
- Roscoe, B. P., K. M. Thayer, K. B. Zeldovich, D. Fushman, and D. N. Bolon, 2014 Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* 425: 1363–1377.
- Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski *et al.*, 2011 An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475: 348–352.
- Sahni, N., S. Yi, Q. Zhong, N. Jaikhani, B. Charlotiaux *et al.*, 2013 Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* 23: 649–657.
- Sanjuan, R., A. Moya, and S. F. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA* 101: 8396–8401.
- Starita, L. M., J. N. Pruneda, R. S. Lo, D. M. Fowler, H. J. Kim *et al.*, 2013 Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. USA* 110: E1263–E1272.
- Thornton, J. W., E. Need, and D. Crews, 2003 Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301: 1714–1717.
- Tokuriki, N., and D. S. Tawfik, 2009a Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* 459: 668–673.
- Tokuriki, N., and D. S. Tawfik, 2009b Protein dynamism and evolvability. *Science* 324: 203–207.
- Tokuriki, N., and D. S. Tawfik, 2009c Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19: 596–604.
- Tsai, I. J., D. Bensasson, A. Burt, and V. Koufopanou, 2008 Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl. Acad. Sci. USA* 105: 4957–4962.
- Villali, J., and D. Kern, 2010 Choreographing an enzyme's dance. *Curr. Opin. Chem. Biol.* 14: 636–643.
- Wagenaar, T. R., L. Ma, B. Roscoe, S. M. Park, D. N. Bolon *et al.*, 2014 Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res* 27: 124–133.
- Walzthoeni, T., A. Leitner, F. Stengel, and R. Aebersold, 2013 Mass spectrometry supported determination of protein complex structure. *Curr. Opin. Struct. Biol.* 23: 252–260.
- Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57–63.
- Weinreich, D. M., N. F. Delaney, M. A. Depristo, and D. L. Hartl, 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312: 111–114.
- Whitehead, T. A., A. Chevalier, Y. Song, C. Dreyfus, S. J. Fleishman *et al.*, 2012 Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* 30: 543–548.
- Wintrobe, P. L., G. I. Makhatadze, and P. L. Privalov, 1994 Thermodynamics of ubiquitin unfolding. *Proteins* 18: 246–253.
- Wright, S., 1932 The roles of mutation, inbreeding, crossbreeding and selection in evolution, pp. 356–366 in *Proceedings of the Sixth International Congress of Genetics*, edited by D. Jones. Ithaca, NY.
- Wylie, C. S., and E. I. Shakhnovich, 2011 A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. USA* 108: 9916–9921.
- Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*. CABIOS 13: 555–556.
- Zuckerandl, E., 1976 Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol.* 7: 167–183.

Communicating editor: J. Rine