CellPress
REVIEWS

Review

Biophysical and Mechanistic Models for Disease-Causing Protein Variants

Amelie Stein,[1,*] Douglas M. Fowler,[2] Rasmus Hartmann-Petersen,[1] and Kresten Lindorff-Larsen [1,*]

The rapid decrease in DNA sequencing cost is revolutionizing medicine and science. In medicine, genome sequencing has revealed millions of missense variants that change protein sequences, yet we only understand the molecular and phenotypic consequences of a small fraction. Within protein science, high-throughput deep mutational scanning experiments enable us to probe thousands of variants in a single, multiplexed experiment. We review efforts that bring together these topics via experimental and computational approaches to determine the consequences of missense variants in proteins. We focus on the role of changes in protein stability as a driver for disease, and how experiments, biophysical models, and computation are providing a framework for understanding and predicting how changes in protein sequence affect cellular protein stability.

The DNA Avalanche and Interpreting Missense Variations

Technological advances in DNA sequencing have made human genome sequencing on a large scale not only feasible but also affordable. The resulting data avalanche has highlighted the challenge of interpreting the phenotypic consequences of genetic variants [1,2]. Variant interpretation is particularly challenging since more than half of the distinct variants found in an analysis of >60 000 human exomes were only observed in a single individual [3] and since many diseases have a complex, polygenic origin [4]. Although the problem is difficult and complicated, the potential to improve the understanding, diagnosis, and treatment of human diseases is enormous (Figure 1).

Missense variants, in which one amino acid is replaced by another, represent more than 40% of the unique variants observed in the Exome Aggregation Consortium database [3]; yet, their phenotypic consequences are often difficult to predict. This is in contrast to nonsense or frameshift variants that cause large changes to the encoded protein and consequently are usually deleterious. As an example, systematic mutagenesis studies of the highly conserved protein ubiquitin have shown that many single amino acid changes only have a minor impact on protein function in a cellular assay [5]. An analysis of similar high-throughput data across multiple proteins suggests that indeed about two-thirds of single amino acid changes have only a minor effect on function [6]. Some variants are, however, severely detrimental and cause essentially complete loss of function. An interesting observation from further studies on ubiquitin is that at least for this highly conserved protein, there can be substantial variation of the effect of a variant depending on the cellular status and conditions, so that most are detrimental under at least one condition [7]. Thus, in a biological and clinical context, there can be wide variability in the number and type of tolerated mutations in a gene [8,9].

In a clinical setting, it would be useful to have robust methods and sufficient data for interpretation of genetic variants and accurate classifications of whether they are pathogenic

Highlights

Human exome sequencing is revealing millions of missense variants that change protein sequences, but their phenotypic consequences are mostly unknown.

Deep mutational scanning and other high-throughput experiments provide simultaneous insights into the effects of thousands of variants.

Loss of protein stability is a common origin of inherited diseases, and computational predictions of protein stability are useful for assessing variant consequences.

Cellular protein quality control provides a mechanistic link between altered protein stability and cellular protein levels and degradation.

Computational biophysics, evolutionary sequence analyses, and machine learning methods each provide information about variant consequences and may potentially be combined.

Mechanistic models for how mutations give rise to disease provide a starting point for therapeutic strategies.

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[2]Departments of Genome Sciences and Bioengineering, University of Washington, Seattle, WA, USA

*Correspondence:
amelie.stein@bio.ku.dk (A. Stein) and
lindorff@bio.ku.dk (K. Lindorff-Larsen).

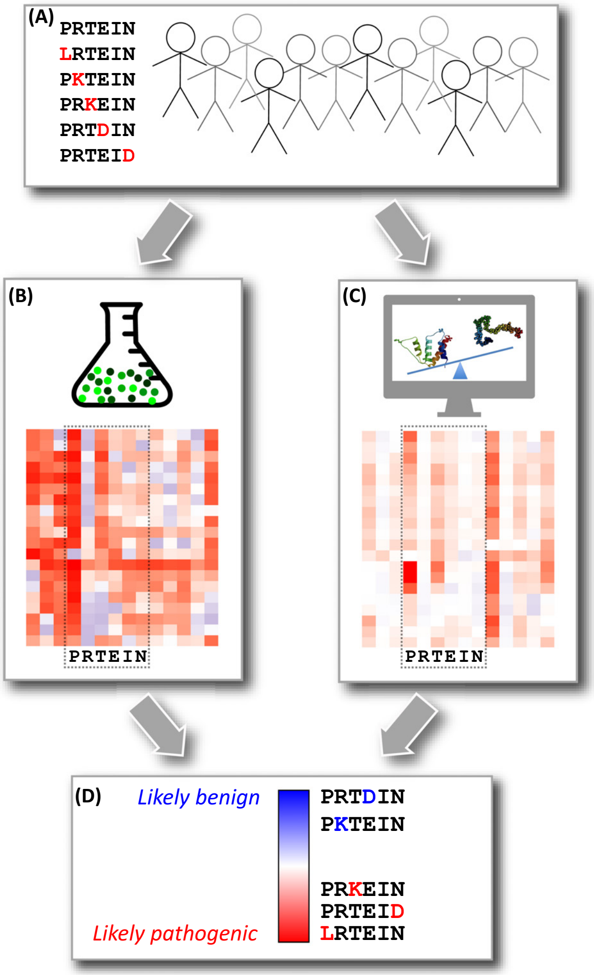


Figure 1. Using High-Throughput Experiments and Computational Methods to Classify Protein Variants. (A) Genomic variation across the human population gives rise to phenotypic variation, some of which is caused by 'protein variants' in which the encoded proteins differ in sequence at a single amino acid position. A key problem is to determine whether such variation has little biological consequence ('benign') or increases the risk for a certain disease ('pathogenic'). (B) High-throughput experiments such as deep mutational scanning (DMS) is one strategy to probe the effect of virtually all possible missense variants in a single experiment and may be summarized by a heatmap that shows whether the variant causes severe loss of function or another property (red), or whether the variant protein behaves similar to the wild type (blue). (C) Alternatively, or as a supplement to experiments, computational methods can be used to predict whether the variant is likely to perturb, for example, activity, stability or other properties important for function. Such prediction methods may for example be based on sequence conservation through evolution, specific biophysical models, or be trained through machine learning to capture experimental observations. (D) The experimental data or computational results are then used, sometimes in concert, to help predict phenotypic consequences of genomic variation of use, for example, in patient diagnosis or gene discovery or to provide mechanistic models of the origins of disease.

Trends in Biochemical Sciences

or benign (Figure 1A) [10]. This is particularly important for diseases for which such information can lead to clinical action [11]. To further our understanding of the origins of disease, it would also be extremely valuable to have reliable predictors of the underlying mechanisms by which variants lead to disease.

There are several conceptual frameworks available to study, model, and predict the phenotypic consequences and pathogenicity of sequence variation. For example, one may use cellular or biochemical assays to quantify the effects of a variant on function and other properties, and recent developments are enabling such studies in high throughput by covering all possible individual amino acid changes (Figure 1B) [12]. Another framework is to use bioinformatics and machine learning methods to integrate existing data, in particular information about sequence conservation, to interpret what sequence variation is compatible with function [13]. Finally, one may use the accumulated knowledge about protein structure, function, and folding to determine the likely effect of a variant (Figure 1C) [14]. These different approaches are not mutually exclusive, and ongoing efforts indeed aim to combine them (Figure 1D).

In this review, we focus on missense variants that result in a change from one amino acid to another (henceforth variants). Furthermore, we focus on recent efforts to understand and predict the effects these variants have on biophysical properties of proteins and consequently their effect on function. While protein-coding regions only make up ~1.5% of the genome, around 5–10% of hits in genome-wide association studies fall into them, although linkage disequilibrium (joint inheritance of elements proximal on a chromosome) makes it challenging to identify precisely which of multiple nearby variants is causal [15]. Beyond diagnosis, we may use existing knowledge of proteins and their cellular pathways to help elucidate the disease-causing mechanisms. Because proteins can be targeted by small molecules or peptides, these insights can potentially open up therapeutic avenues.

Loss of Protein Stability as Origin of Disease

Protein stability is one of the most basic properties of a protein and may be strongly affected by missense variants. As most proteins need to be folded to function, loss of stability may lead to loss of function. In the context of a biophysical or biochemical experiment, stability generally refers to the thermodynamic or kinetic stability between a fully folded and globally unfolded state, but in a cellular and disease context many other factors and protein conformations play a role. These factors include interactions with the cellular protein quality control system, protein–protein interactions, cellular trafficking, and post-translational modifications. Analyses linking the predicted effect of amino acid changes to the thermodynamic stability of a protein with its cellular stability and pathogenicity suggest that loss of stability could be a main driver and origin of inherited diseases [16–20]. Thus, an improved understanding of the complex relationship between protein sequence, structure, folding, and cellular stability could provide new possibilities for diagnosis and even treatment.

Experimental studies of protein folding and stability *in vitro* and *in vivo* may provide detailed, quantitative descriptions and mechanistic insights of the effects of individual amino acid changes. Until recently, they were limited to studying the effects of a few variants, generally limiting studies to retrospective analyses of variants already seen in patients. Recent developments in high-throughput experiments are, however, beginning to provide us with orders of magnitude more data to improve our models and understanding of protein stability and to perform prospective studies of variants not yet seen in patients [21]. By leveraging the same advances in DNA sequencing that are enabling cheap sequencing of human genomes, high-throughput experiments are making it possible to study the effects of sequence variation on a scale not previously possible [22]. Combined with genetic selection systems, DNA sequencing methods can also be used to study the mechanisms and sequence specificity of cellular protein quality control [23].

Together, these developments are now being put to use to improve the predictions of clinical outcomes and to provide mechanistic models for diseases. Below, we review recent developments in these areas, focusing on the role that loss of protein stability and resulting loss of function play in human diseases. We begin with an overview of the cellular protein quality control system that recognizes unstable or misfolded proteins and targets them for degradation and thus is the mechanistic link between loss of stability and decreased cellular abundancy of proteins. We proceed to show how deep mutational scanning (DMS) experiments are transforming our ability to study functional and mechanistic consequences of variants. We then describe recent developments in using computational methods to predict the consequences of variants and end by outlining how insights into the mechanisms underlying loss of cellular protein stability may be used to develop new therapies.

Cellular Protein Quality Control

Since structurally destabilized or misfolded proteins may form various toxic inclusions or aggregates, all organisms have evolved several protective measures to guard against these potentially harmful proteins. Collectively, these mechanisms are known as protein quality control (PQC) systems, with the two main strategies being either refolding or degradation of the misfolded proteins [24,25].

During or after synthesis, proteins may undergo transitions through various metastable folding intermediates towards the native state and be protected from aggregation by molecular chaperones; in a similar manner chaperones may also catalyze the refolding of proteins that become damaged after synthesis [24]. Degradative PQC, in contrast, relies on proteases to irreversibly clear the intracellular environment of non-native proteins. Both of these PQC systems must be highly specific for incorrectly folded proteins but also broadly inclusive to ensure that many structurally diverse proteins can be targeted. Accordingly, defects in either of these systems can lead to accumulation of toxic protein species that, in turn, may trigger diseases, including several neurodegenerative disorders [26,27]. Conversely, an overaggressive destruction of structurally destabilized, but functional, proteins has been linked to various hereditary diseases, including cystic fibrosis [28,29] and Lynch syndrome [19,30,31]. It therefore becomes clear that substrate selection is a trade-off between specificity and recognition of a wide variety of substrates.

In eukaryotes, most protein degradation occurs in the cytosol and nucleus via the ubiquitin-proteasome system (UPS) or the autophagy-lysosomal pathway [32], with the latter system typically responsible for the degradation of highly misfolded and insoluble protein aggregates. Aggregation has also been linked to several diseases; however, this is beyond the scope of this review article and is reviewed in [33]. The UPS generally targets soluble or partially soluble proteins through a process involving conjugation of a polyubiquitin chain to the substrate protein, thus targeting it to degradation by the 26S proteasome. Ubiquitin conjugation is catalyzed by an enzymatic cascade that includes substrate-specific E3 ubiquitin-protein ligases that add the ubiquitin chains to the target protein. The discriminating feature in a destabilized protein that elicits its recognition by E3s and degradation, the so-called degron, is despite tremendous recent efforts [23,34–36], not completely understood, but it is likely to involve hydrophobic regions that are buried in the native protein but exposed in misfolded proteins (Figure 2) as well as intrinsically disordered segments where degradation can be initiated. We refer the reader to recent reviews of the role and components of the PQC that are important to the degradation of misfolded proteins and of the molecular and biophysical origins of proteasomal degradation [37–39].

In the context of disease-causing variants, a key question is how much structural destabilization is tolerated before the PQC system kicks in? Recently, it was shown that the degree of protein destabilization correlates with the turnover rate in the Lynch-syndrome related protein MSH2 [19]. Surprisingly, however, as little as 3 kcal mol^{-1} was sufficient to trigger degradation [19]. Although this figure is likely to vary from protein to protein, depending on how stable the wild-type protein is, a 3-kcal mol^{-1} destabilization is certainly not dramatic, compared with, for example, the average stability of 5 kcal mol^{-1} for a series of small proteins [40]. It is, however, in agreement with genetic studies in yeast that have shown that the PQC system operates by following a better-safe-than-sorry principle and is thus highly diligent and prone to target proteins that are only slightly perturbed and still functional [31,41,42].

A key problem to tackle in the future is to understand better what structural features are actually recognized by the PQC system and thus refine our understanding of degradation signals, both

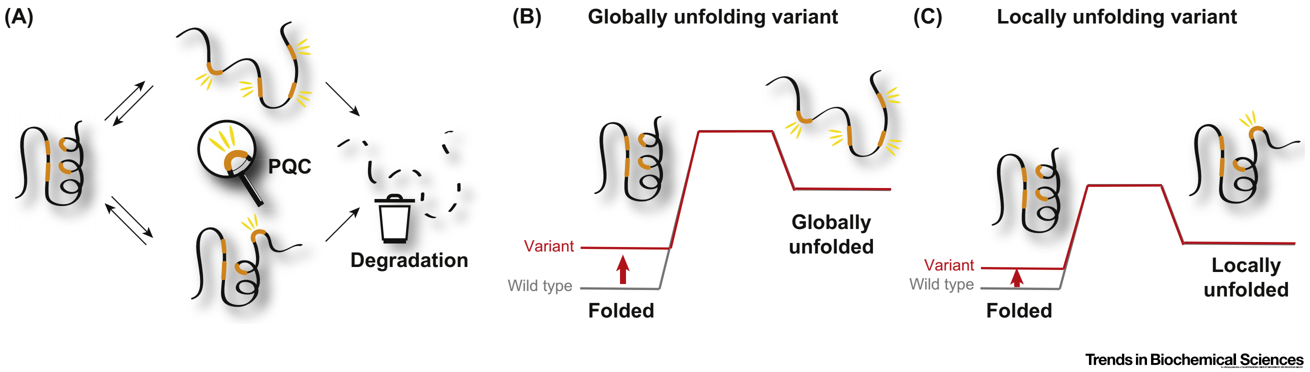


Figure 2. Mechanisms for Cellular Protein Quality Control and Degradation and Effects of Sequence Variation on the Folding Energy Landscape. (A) In a folded protein (left), the degradation signals (degrons, orange) are generally buried inside the protein. Upon local and partial unfolding (bottom route) or full unfolding (top route), one or more degrons may become exposed. The cellular protein quality control (PQC) components (magnifying glass), such as molecular chaperones and E3 ubiquitin-protein ligases, scan the cell for such degradation signals and target the substrates for degradation (right). Variants may affect all of these steps including increasing the populations of unfolded or partially unfolded states, or creating or removing degron sequences. (B) A globally destabilizing variant brings the free energy of the folded conformation closer to that of the fully unfolded state, increasing the population of this state and making the protein more easily targeted for degradation. (C) Because local unfolding involves smaller free energy differences, amino acid changes with more modestly destabilizing effects may still cause a substantial increase in locally unfolded states, and possible exposure of degrons. In this way such variants can have a stronger effect in the cell than one would expect from the predicted thermodynamic change of global stability.

at the stage of ubiquitin conjugation and when substrates are degraded at the proteasome [34,35,37–39]. For example, it is unclear whether cells generally recognize global or local unfolding events, and what the relationship is between such unfolding events and transient exposure of degron sequences (Figure 2). In this context, a single amino acid change causing a destabilization of a few kilocalories per mole could cause a substantial increase in the population of locally unfolded structures that, in turn, would lead to degradation and insufficient levels of the affected protein.

Deep Mutational Scanning

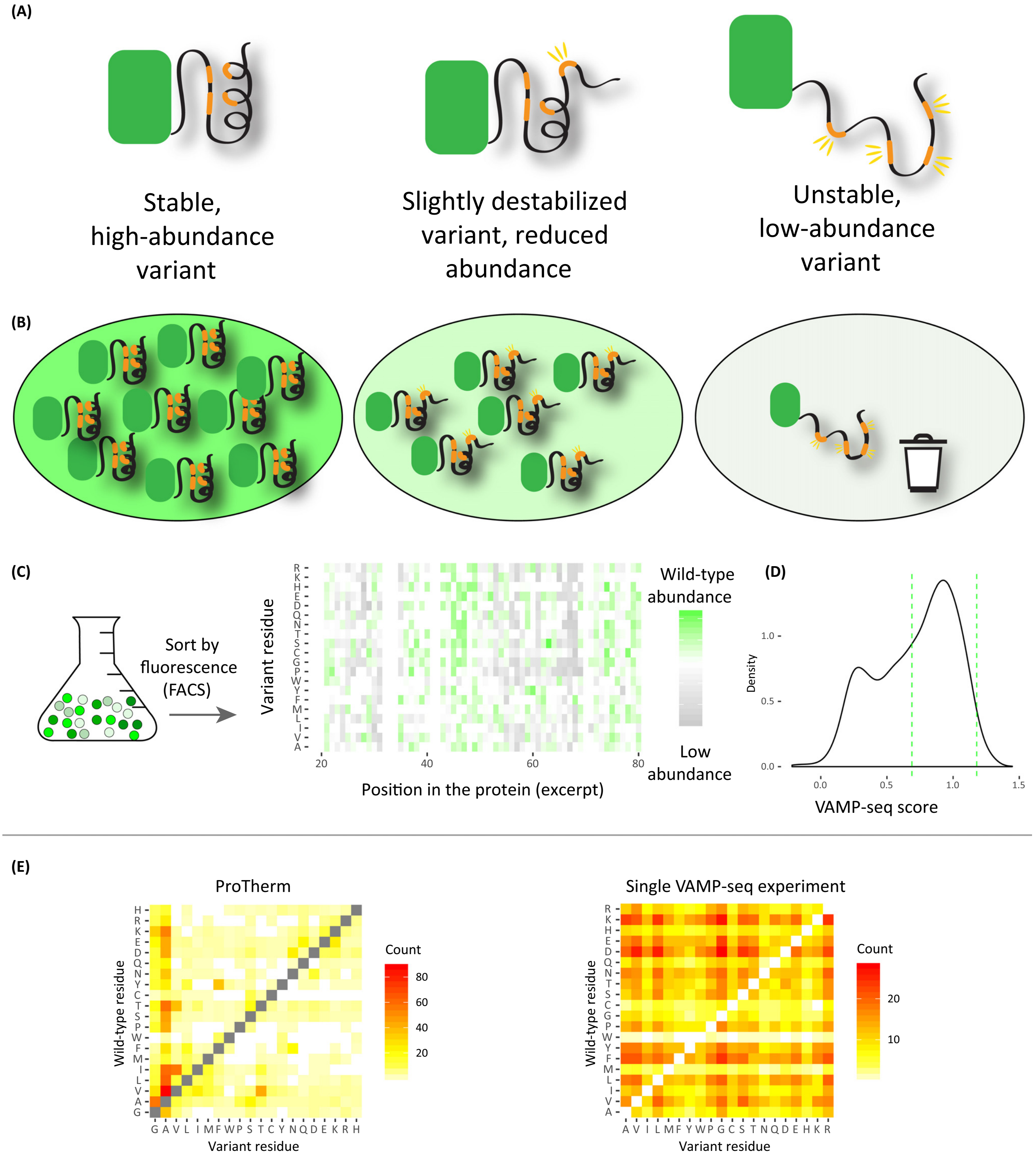
Much of what we know about how proteins fold and are stabilized has been learned by studying individual amino acid changes. However, this one-at-a-time approach probes only a tiny fraction of the possible genetic variation we could observe in an individual and hence limits our understanding and ability to predict phenotypic consequences. DMS experiments leverage cheap DNA sequencing to probe the effects of hundreds or thousands of variants in a single, multiplexed assay [22,43]. First, selection for a protein property of interest is applied to a large library of variants. Selections used so far include coupling protein activity to cell growth, coupling protein activity or stability to a fluorescent reporter, or selecting for ligand binding using phage or yeast display. Variants in the library change in frequency depending on how well they able to perform under selective conditions. Finally, the frequency of each variant before and after the selection is read out using next-generation DNA sequencing, and each variant's change in frequency is used to compute a score that quantifies the effect of the variant on the property and conditions selected under.

Most applications of DMS have used selection for a biological function of the protein that can be probed in high throughput. For example, in a recent tour de force, the effect of variants of the *BRCA1* gene were assayed using saturation genome editing. Here, approximately 4000 variants were introduced into 13 of *BRCA1*'s 24 exons using CRISPR/Cas9 editing of the genomic copy of *BRCA1* in a haploid cell line [12]. The functional consequences of each variant on cell viability were measured using next-generation sequencing and correlated strongly with existing expert-based assessment of pathogenicity. Variants that are common in the human population were more likely to be scored as functional in the assay. Importantly, this experiment also provided functional data for the several thousand variants that have not yet been seen in any patient. These unseen variants are of unknown pathogenicity, so the functional data will be of immediate use if any of them are seen in the future. An interesting observation was also that ~90% of all loss-of-function variants had no substantial changes in mRNA levels, suggesting that most missense variants, at least in *BRCA1*, affect function at the protein level. As observed from the results on ubiquitin discussed in the introduction, as well as a dual-assay DMS study of BRCA1 [44], different assays and conditions might reveal different sensitivities to variants. Furthermore, in a physiological context, individual genes vary in how much they tolerate sequence variations [8,9], making it important to determine the relationship between the DMS results and pathogenicity.

The results of growth-based saturation genome editing experiments like those described for *BRCA1* above depend on the combined effects that sequence variation may have on numerous properties including RNA splicing, expression levels, protein function, protein–protein interaction, post-translational modifications, and protein folding and stability. Because the cellular growth rate may capture many of the biologically relevant effects of variants it can be extremely accurate and useful for assessing the pathogenicity. On the other hand the results may, however, be less informative for disentangling the mechanism by which each variant exerts an effect, and the knowledge obtained is not easily transferable to studying the effects of variants in other proteins.

To enable more widespread analysis of variant consequences without needing to establish protein-specific assays and to learn more general rules regarding the relationship between protein stability and cellular abundance, we have recently developed variant abundance by massively parallel sequencing (VAMP-seq; Figure 3). VAMP-seq measures the impact of variants on the steady-state cellular abundance of a protein [45]. Here, a library of variants of the protein of interest is fused to GFP (Figure 3A). Then, the library is expressed in cultured mammalian cells such that each cell expresses one and only one variant (Figure 3B). The stability of the variant dictates the stability of the GFP fusion, so each cell's GFP fluorescence reports on the abundance of the protein variant. Cells are sorted into bins based on their fluorescence, next-generation sequencing is used to determine the frequency of every variant in each bin, and variant frequencies are used to compute abundance scores (Figure 3C). Thus, a single VAMP-seq experiment provides quantitative abundance data for thousands of variants simultaneously and enables one to separate variants with modest effects on stability from those that are substantially destabilizing (Figure 3D).

In the context of enabling computational prediction methods, it is worth highlighting that a single VAMP-seq experiment provides information about a number of variants comparable in size to the entire database used to train current state-of-the-art models for predicting protein stability [46,47] (Figure 3E). Another advantage of VAMP-seq and other DMS experiments is that they often target most or all of the 19 possible amino acid substitutions at each position. Thus, unlike the majority of available biophysical data that are highly biased [48] and mostly consist of side chain truncations to alanine or glycine (Figure 3E), comprehensive stability data can be used to guide the development of improved prediction methods. While most biophysical experiments probe only a subset of the possible mutations, a recent study made it possible to examine how accurate DMS experiments can probe protein stability [49]. By comparing *in vitro* measurements of thermodynamic stability, performed in high throughput, with the results of a mechanistic analysis of a DMS experiment, the authors showed that DMS may indeed provide quantitative data on protein stability.


(A)
Stable, high-abundance variant
Slightly destabilized variant, reduced abundance
Unstable, low-abundance variant
(B)
(C)
Sort by fluorescence (FACS)
Variant residue
R K H E D Q N T S C G P W Y F M L I V A
20 40 60 80
Position in the protein (excerpt)
Wild-type abundance
Low abundance
(D)
Density
0.0 0.5 1.0
0.0 0.5 1.0 1.5
VAMP-seq score
(E)
ProTherm
Wild-type residue
H R K E D Q N Y C T S P W F M I L V A G
G A V L I M F W P S T C Y N Q D E K R H
Variant residue
Count
80 60 40 20
Single VAMP-seq experiment
Wild-type residue
R K H E D Q N T S C G P W Y F M L I V A
A V I L M F Y W P G C S T N Q D E H K R
Variant residue
Count
20 10

Trends in Biochemical Sciences

Figure 3. Deep Mutational Scanning for Protein Stability and Variant Abundance. (A–C) Outline of the variant abundance by the massively parallel sequencing (VAMP-seq) method [45]: (A) generation of a large library of variants, typically all possible 19 variants at each site, and fusion to GFP; (B) abundance of the respective variant fusion construct determines each cell's fluorescence; and (C) fluorescence-activated cell sorting (FACS) followed by sequencing and data analysis allows for the *(Figure legend continued on the bottom of the next page.)*

DMS is already a widely applied method and will become even more useful as methods for generating and sequencing variant libraries improve and decrease in cost. The resulting data are useful in the clinic because they can be used to aid the interpretation of any variant including those not yet seen in any patient and because they may be used to train better prediction methods. We also note that DMS and related high-throughput experiments may provide very useful information for understanding and improving protein function and stability for example in protein engineering and design [50,51].

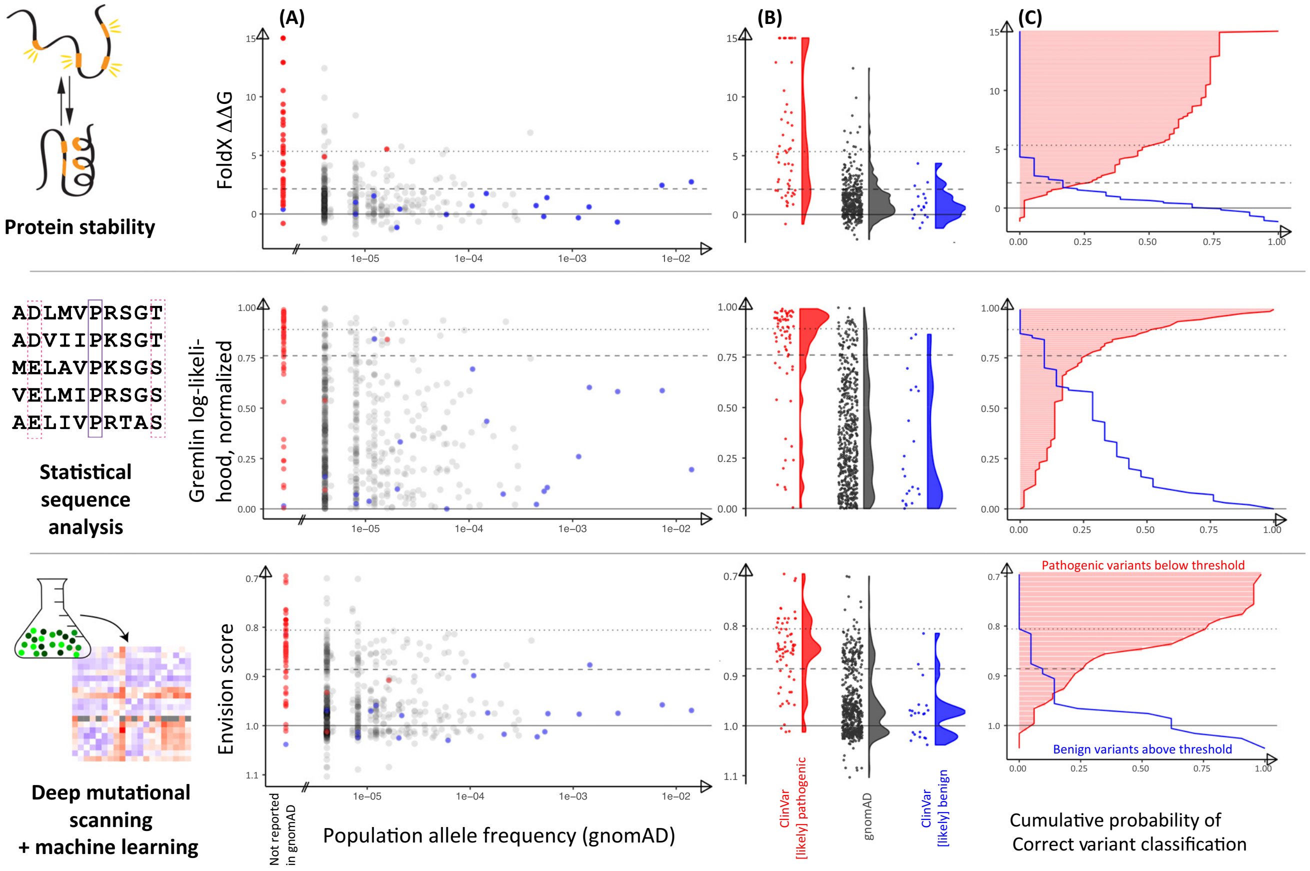
Predicting the Consequences of Missense Variation

While experimental testing of variants is expanding in scope and scale, computational predictions of variant consequences will continue to be the only widely applicable method to assess pathogenicity for the foreseeable future. Several predictors have been trained specifically for this purpose, often using known benign and pathogenic variants [52]. Here, we instead focus on three distinct approaches developed to address more general questions concerning how changes in the protein sequence affect, for example, protein stability or general functional properties. These methods have not been specifically trained on pathogenic variants; instead, they were created to capture thermostability of folding, evolutionary tolerance, and patterns observed in DMS experiments, respectively. To illustrate the outcome and performance of these three classes of prediction methods, we show the results of stability calculations (Figure 4A), a sequence likelihood model (Figure 4B), and the DMS-based prediction method (Figure 4C) on the protein MSH2 and discuss them in more detail below.

Modeling amino acid substitution(s) directly in a protein's 3D structure should, in principle, enable an accurate assessment of the resulting change in folding energy. Two tools that take this approach are FoldX [47] and Rosetta [53]; each predict the effect of an amino acid change on stability with an accuracy of about 1 kcal mol^{-1} and a correlation coefficient of ~0.7 (depending on test set; [46]). In addition to predicting stability effects, these and related methods have been shown to successfully identify pathogenic variants in several proteins [16,19]. In selected cases, experimental validation yielded a correlation between the predicted loss of stability and cellular protein levels [19,45,54]. In addition to classifying unstable variants as pathogenic, stability predictions have the advantage of indicating the likely underlying mechanism; this information is useful when developing therapeutic strategies (see below).

Prediction methods that focus on a specific mechanism such as loss of stability will, of course, not capture variants that give rise to disease via different mechanisms. Thus, stability predictions are most useful when combined with other predictors [52,55–57]. Analysis of the conservation patterns in a multiple sequence alignment of a protein family is a powerful and general approach to identify substitutions that are pathogenic by their paucity in, or absence from, the alignment and indeed is used in most prediction methods [52]. One caveat with analyzing the conservation at individual sites is that it neglects the context in which the variation occurs although such effects may be important [58]. A recent development is, however, the construction of higher-order statistical models that examine both conservation at individual sites and also between multiple sites [59–62]. While these latter approaches generally provide greater accuracy than methods that analyze each site independently [63], they require a larger

quantification of the abundance of each variant. (D) Distribution of VAMP-seq scores for missense variants in the protein PTEN, normalized such that a value of one corresponds to the wild-type protein sequence and zero to the average of the 1% lowest scoring variants [45]. Green lines indicate the 5th and 95th percentile for synonymous variants; 56% of the missense variants fall within this range. (E) Accurate biophysical measurements of the change in protein stability upon amino acid changes have been collected over many years [46] but are dominated by substitutions to alanine, and a few other chemically, structurally, or biophysically motivated substitutions [82] (left). By contrast, a single VAMP-seq experiment provides data for a comparable number of variants but is less biased chemically (right).



Trends in Biochemical Sciences

Figure 4. Three Paradigms for Predicting the Consequences of Amino Acid Changes. We illustrate the utility of stability predictions (top), evolutionary analyses (middle), and a regression model trained on deep mutational scanning data (bottom) to predict the consequences for pathogenic and benign MSH2 variants from the ClinVar database [68]. (A) The allele frequencies in the gnomAD database of genome sequences (gnomad.broadinstitute.org) are plotted against the predicted score of the variant. The variant scores are ordered so that detrimental variants are shown at the top, and stability prediction scores were truncated at 15 kcal mol^{-1}. Red and blue points are those reported as (likely) pathogenic and benign, respectively, in ClinVar. The left-most 'column' of points (labeled 'not reported in gnomAD') contains variants reported in ClinVar, but not observed in gnomAD; they mostly correspond to known pathogenic variants expected to be found at very low allele frequencies. (B) Raincloud plots [83] illustrating the predicted score distributions of pathogenic (red), population (gray), and benign (blue) variants. For all three prediction methods there is a clear, yet also non-perfect, separation between pathogenic and benign variants. (C) Cumulative distribution functions showing which fraction of variants are above/below any given score threshold. The red curve shows the fraction of pathogenic variants below the value (false negatives), and the blue curve the fraction of benign variants above the threshold (false positives). The horizontal dashed lines indicate the respective threshold for 25% false negative predictions, and the dotted lines are the thresholds for no false positives. Solid lines indicate the respective predictor's value for the wild type. Overall, the plots illustrate that all three predictors correctly identify many of the pathogenic variants as detrimental, and most of the benign variants as tolerated. The 'area under the curve' (AUC) in a receiver operating characteristic (ROC) analysis is 0.91, 0.90, and 0.91 for the three methods, respectively. To address the imbalance between the sizes in the pathogenic and benign datasets, the pathogenic dataset was split in three; these AUCs are averages over these three ROC analyses.

number of homologous sequences [64]. This restriction arises because the methods involve building global sequence models rather than examining each site independently. Analyses that consider both site conservation and pairwise co-varying positions have successfully been applied to predict variant pathogenicity [63,65] and more recently, more general models have been introduced [13]. Co-variation also occurs between different genes, and may be indicative of direct protein–protein interactions.

Because evolutionary conservation across a protein family is likely to capture residues required for the protein's core function, these approaches can identify variants that affect many protein properties including stability, enzymatic activity, post-translational modifications, or protein–protein interactions. Thus, a conserved variant may be neutral from the perspective of thermodynamic folding energy but have strong functional consequences. By contrast, evolutionary sequence analysis may miss pathogenic changes where the residue in question is critical only for human biology, or in a small branch of the protein family's phylogenetic tree, or where the variation has specific effects on regulation or modifications. In this context, recent analyses focusing on sequence conservation in non-human primates are particularly interesting [66].

As an alternative to analyses of conservation through deep multiple sequence alignments, one may use other sources of data to learn what kind of amino acid changes typically lead to perturbed function. Here, DMS experiments now provide us with a large collection of the functional effects of tens of thousands of substitutions across a diverse set of proteins [6]. Annotation of this functional data with biochemical and coarse-grained structural features was combined with machine learning to create Envision, a tool for quantitative prediction of the effect of missense variants [67]. In contrast to the biophysical modeling and sequence conservation analysis approaches discussed above, Envision does not require specific data on the protein in question beyond its sequence and is thus more widely applicable than stability calculations and statistical sequence analysis, yet it successfully identified many pathogenic variants in a recent benchmark [67].

As an example of the power of using these three prediction paradigms, we show their application to the protein MSH2, where variants may lead to cancer predisposition (Lynch syndrome) (Figure 4). Specifically, as described previously [19], we used FoldX [47] to calculate changes in protein stability from the structure of MSH2. We also used Gremlin [60], a method to learn a probabilistic model from a multiple sequence alignment, to analyze conservation patterns in MSH2 [19]. Finally, we used a Envision [67], the above-mentioned machine learning method trained on DMS data, and structure and sequence features to predict the consequences of variants. In contrast to our previous work that focused on a smaller set of variants, we here used ClinVar [68], an archive of medically important variants and phenotypes, to select 66 pathogenic and 21 benign variants, and we also analyzed the 587 missense variants of MSH2 found in the Genome Aggregation Database (gnomAD) [3], which aggregates information from several different exome and whole-genome sequencing projects.

The results show clearly that although these methods have not been trained on population genetics data or disease variants, they are able to separate known disease-causing variants from benign variants with relatively high accuracy. For example, benign variants generally have modest effects on stability, whereas many pathogenic variants are highly destabilizing. It is also worth noting that only 3 of the 66 pathogenic variants seen in ClinVar have actually been observed in the ~150 000 genome and exome sequences available in gnomAD, while 19 of 21 benign variants have been observed. Thus, there is a clear trend that more common population variants are predicted to have milder effects, whereas many uncommon variants and pathogenic variants are predicted to have more dramatic effects (Figure 4A). These observations imply that there is a clear difference in the distribution of predicted scores between benign and pathogenic variants (Figure 4B) that, in turn, can be transformed into relatively accurate predictions (Figure 4C). Nonetheless, the analyses also show that these predictions of functional effects are not yet alone sufficient to fully separate benign from pathogenic variation. Thus, we note that many other methods exist for predicting pathogenicity. We have chosen three methods that each aim to predict rather basic properties, are not based on analysis of any known disease-causing variants, and that are generally applicable to a wide range of proteins.

In a clinical setting and for optimal prediction accuracy one could combine assessment of the effects of variants on multiple protein properties [52,55–57,69,70].

Therapeutic Possibilities

In addition to the prospect for improved diagnosis via prediction of pathogenicity, the experimental and computational studies discussed above provide new opportunities for treatment of diseases. For variants that give rise to disease via loss of stability, intracellular degradation and thereby loss of function, it might be possible to rescue function via restabilization. In particular, because the PQC is overzealous in targeting potentially functional, but mildly destabilized proteins, many disease-causing variants might be sufficiently functional that pathogenicity could potentially be averted if the proteins were stabilized [31] (Figure 5).

The most dramatic approach is perhaps to inhibit the proteasome, and proteasome inhibitors are indeed already approved drugs [71]. In many cases, a more direct and elegant approach might be to target the components in the PQC that are relevant for degrading a specific disease-causing variant. To enable this approach, we need to map in much greater detail the E3 enzymes and chaperones involved in recognizing specific substrates and targeting them for degradation. As an example, in yeast, certain mutant variants of MSH2 linked to Lynch syndrome can be rescued by deleting the E3 ligase that targets the MSH2 variants for degradation, thus restoring cellular MSH2 protein levels and MSH2 function [30]. Thus, targeting the equivalent but still unknown [72] human E3 ligase may provide treatment options for individuals with certain MSH2 variants. Since several PQC E3s display overlapping substrate specificity [73], this will likely be complicated. Other strategies involve increasing or decreasing the levels of chaperones that either aid in refolding or degradation [74,75].

Some protein variants might be so unstable that even inhibiting their degradation would not be sufficient to restore cellular stability and function. These variants might, however, be rescued via small molecules that bind directly to the destabilized variant protein [76]. This chemical chaperone or corrector approach has already been shown to rescue function for example in mutant TP53 [77] and CFTR [78].

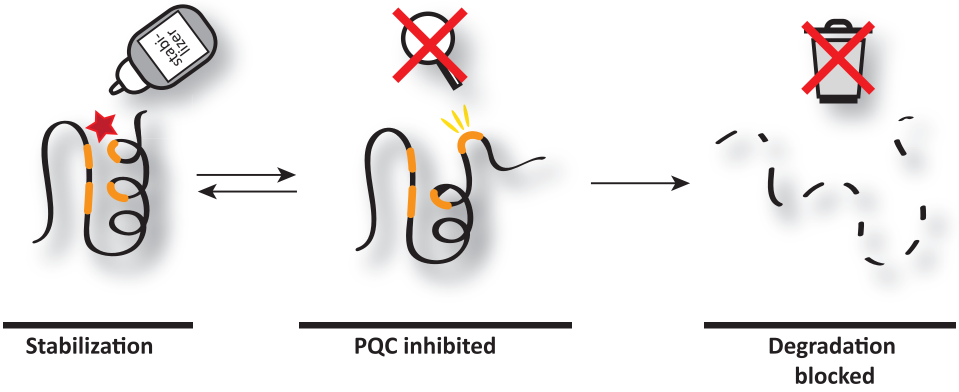


Figure 5. Rescuing Protein Stability as a Strategy for Therapy. The cellular levels of a destabilized protein variant may be increased by blocking the protein quality control (PQC) system (magnifying glass; middle) or the degradation machinery (trashcan; right). Alternatively, a small molecule (star) that associates with the native form of the protein may act to stabilize the protein.

Concluding Remarks

Widespread access to cheap DNA sequencing is transforming medicine and science. Within precision medicine, genome or exome sequencing provides possibilities for finding causal variants and for improved diagnosis and possible treatment. Within protein science, DMS experiments are enabling the study of the effects of thousands of variants in a single experiment. Recent efforts are bringing these fields together by using DMS to help classify variants as benign or pathogenic and by providing data to benchmark or train prediction methods for variant classification. These approaches may be particularly important for so-called rare genetic diseases that are difficult to diagnose from population-based studies [79]. Accurate predictions of variant consequences may also be useful in finding rare causal variants or aggregating information across variants.

So far, these approaches have mostly been applied to simple, monogenic Mendelian disorders. In the future, it will be interesting to investigate whether they can improve polygenic risk scores [80] that aggregate information across variants in multiple genes (see Outstanding Questions). Here, it is worth noting how stability predictions for protein–protein complexes provide a direct mechanism for finding apparently non-additive effects. For example, two variants that individually only cause a mild change in the stability of the complex may, when combined, have a dramatic effect because of the non-linear relationship between energy and population of the complex.

One of the problems in assessing the importance of loss of stability for disease is that we do not fully understand when and why the current prediction methods fail. This is, in part, due to the fact that they were trained and benchmarked on a biased dataset that mostly contains experiments where a large amino acid is mutated to a smaller amino acid, often alanine or glycine. We expect that unbiased data from DMS experiments will be extremely useful in assessing and parameterizing prediction methods for a much wider set of amino acid changes. Stability calculations generally predict the consequence on the global thermodynamic stability describing the equilibrium between the fully folded state and an implicitly represented fully unfolded state. Such calculations are thus not directly applicable to modeling effects on local unfolding events, although as described above these could play a central role in cellular stability, and we need better and computationally efficient tools to study these. An important problem to tackle in the future is also to map genetic variants to accurate structural models for the entire human proteome [81] and to develop prediction methods that better use structural information and are robust towards structural noise in homology models. Finally, an important open question is how the different prediction methods are best combined and how they can both provide accurate predictions of pathogenicity and aid in developing mechanistic hypotheses for the origin of disease.

Outstanding Questions

What are the structural features of the unfolded and misfolded states, and how are they recognized by the PQC system?

Are there generic PQC components including chaperones and E3s that target a wide range of human missense variants?

When current predictors fail, why is that? Can we develop confidence scores to identify less reliable predictions?

Can biophysics, statistical sequence analysis and machine learning on DMS data improve polygenic risk scores?

How are predictors best combined, both to improve accuracy and to develop mechanistic hypotheses for the origin of genetic diseases?

Can we develop therapeutic strategies to target many different variants in a single protein, or variants in different proteins that are degraded by similar pathways?

Acknowledgments

We thank Drs Sofie V. Nielsen and Caspar Elo Christensen for helpful comments on the manuscript and assistance with preparing the figures. Our work in this area has been supported by grants from the Lundbeck Foundation (A.S., R.H-P., and K.L-L.), The Danish Cancer Society (R.H-P.), the Novo Nordisk Foundation (R.H-P. and K.L-L.), the Danish Council for Independent Research (Natural Sciences) (to R.H-P.), and the National Institute of General Medical Sciences (1R01GM109110 to D.M.F.). D.M.F. is a CIFAR Azrieli Global Scholar.

References

1. Shendure, J. and Akey, J.M. (2015) The origins, determinants, and consequences of human mutations. *Science* 349, 1478–1483
2. Manolio, T.A. *et al.* (2017) Bedside back to bench: building bridges between basic and clinical genomic research. *Cell* 169, 6–12
3. Lek, M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291
4. Martin, H.C. *et al.* (2018) Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 42, 1161–1164

5. Roscoe, B.P. *et al.* (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* 425, 1363–1377
6. Gray, V.E. *et al.* (2017) Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* 207, 53–61
7. Mavor, D. *et al.* (2018) Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biol. Open* 7, bio036103–8
8. Bartha, I. *et al.* (2018) Human gene essentiality. *Nat. Rev. Genet.* 19, 51–62
9. Schaafsma, G.C.P. and Vihinen, M. (2017) Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. *Hum. Mutat.* 38, 839–848
10. Richards, S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424
11. MacArthur, D.G. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–476
12. Findlay, G.M. *et al.* (2018) Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217–222
13. Riesselman, A.J. *et al.* (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 15, 816–822
14. Kroncke, B.M. *et al.* (2016) Documentation of an imperative to improve methods for predicting membrane protein stability. *Biochemistry* 55, 5002–5009
15. Brodie, A. *et al.* (2016) How far from the SNP may the causative genes be? *Nucleic Acids Res.* 44, 6046–6054
16. Pey, A.L. *et al.* (2007) Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. *Am. J. Hum. Genet.* 81, 1006–1024
17. Casadio, R. *et al.* (2011) Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* 32, 1161–1170
18. Pal, L.R. and Moult, J. (2015) Genetic basis of common human disease: insight into the role of missense SNPs from genome-wide association studies. *J. Mol. Biol.* 427, 2271–2289
19. Nielsen, S.V. *et al.* (2017) Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet.* 13, e1006739
20. Redler, R.L. *et al.* (2016) Protein destabilization as a common factor in diverse inherited disorders. *J. Mol. Evol.* 82, 11–16
21. Starita, L.M. *et al.* (2017) Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* 101, 315–325
22. Fowler, D.M. and Fields, S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801–807
23. Geffen, Y. *et al.* (2016) Mapping the landscape of a eukaryotic degronome. *Mol. Cell* 63, 1055–1065
24. Hartl, F.U. *et al.* (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475, 324–332
25. Kettern, N. *et al.* (2010) Chaperone-assisted degradation: multiple paths to destruction. *Biol. Chem.* 391, 481–489
26. Ciechanover, A. and Kwon, Y.T. (2017) Protein quality control by molecular chaperones in neurodegeneration. *Front. Neurosci.* 11, 185
27. Rubinsztein, D.C. (2006) The roles of intracellular protein-degradation pathways in neurodegeneration. *Nature* 443, 780–786
28. Ahner, A. *et al.* (2007) Small heat-shock proteins select deltaF508-CFTR for endoplasmic reticulum-associated degradation. *Mol. Biol. Cell* 18, 806–814
29. Meacham, G.C. *et al.* (2001) The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. *Nat. Cell Biol.* 3, 100–105
30. Arlow, T. *et al.* (2013) Proteasome inhibition rescues clinically significant unstable variants of the mismatch repair protein Msh2. *Proc. Natl. Acad. Sci. U. S. A.* 110, 246–251
31. Kampmeyer, C. *et al.* (2017) Blocking protein quality control to counter hereditary cancers. *Genes Chromosomes Cancer* 56, 823–831
32. Kwon, Y.T. and Ciechanover, A. (2017) The ubiquitin code in the ubiquitin-proteasome system and autophagy. *Trends Biochem. Sci.* 42, 873–886
33. Chiti, F. and Dobson, C.M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* 86, 27–68
34. Maurer, M.J. *et al.* (2016) Degradation signals for ubiquitin-proteasome dependent cytosolic protein quality control (CytoQC) in yeast. *G3 (Bethesda)* 6, 1853–1866
35. Rosenbaum, J.C. *et al.* (2011) Disorder targets misorder in nuclear quality control degradation: a disordered ubiquitin ligase directly recognizes its misfolded substrates. *Mol. Cell* 41, 93–106
36. van der Lee, R. *et al.* (2014) Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep.* 8, 1832–1844
37. Yu, H. and Matouschek, A. (2017) Recognition of client proteins by the proteasome. *Annu. Rev. Biophys.* 46, 149–173
38. Enam, C. *et al.* (2018) Protein quality control degradation in the nucleus. *Annu. Rev. Biochem.* 87, 725–749
39. Clausen, L. *et al.* (2019) Protein stability and degradation in health and disease. *Adv. Protein Chem. Struct. Biol.* 114, 61–83
40. Maxwell, K.L. *et al.* (2005) Protein folding: defining a standard set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* 14, 602–616
41. Gardner, R.G. *et al.* (2005) Degradation-mediated protein quality control in the nucleus. *Cell* 120, 803–815
42. Kriegenburg, F. *et al.* (2014) A chaperone-assisted degradation pathway targets kinetochore proteins to ensure genome stability. *PLoS Genet.* 10, e1004140
43. Fowler, D.M. *et al.* (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* 9, 2267–2284
44. Starita, L.M. *et al.* (2015) Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200, 413–422
45. Matreyek, K.A. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50, 874–882
46. Ó Conchúir, S. *et al.* (2015) A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. *PLoS One* 10, e0130433–18
47. Guerois, R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387
48. Yang, Y. *et al.* (2018) PON-tstab: protein variant stability predictor: Importance of training data quality. *IJMS* 19, 1009
49. Nisthal, A. *et al.* (2018) Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *bioRxiv* http://dx.doi.org/10.1101/484949
50. Wrenbeck, E.E. *et al.* (2016) Deep sequencing methods for protein engineering and design. *Curr. Opin. Struct. Biol.* 45, 36–44
51. Gupta, K. and Varadarajan, R. (2018) Insights into protein structure, stability and function from saturation mutagenesis. *Curr. Opin. Struct. Biol.* 50, 117–125
52. Niroula, A. and Vihinen, M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.* 37, 579–597
53. Park, H. *et al.* (2016) Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212
54. Bershtein, S. *et al.* (2013) Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol. Cell* 49, 133–144
55. Raimondi, D. *et al.* (2018) Large-scale in-silico statistical mutagenesis analysis sheds light on the deleteriousness landscape of the human proteome. *Sci. Rep.* 8, 16980

56. De Baets, G. *et al.* (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* 40, D935–D939

57. Wagih, O. *et al.* (2018) A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* 14, e8430

58. Jordan, D.M. *et al.* (2015) Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* 524, 225–229

59. Weigt, M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* 106, 67–72

60. Balakrishnan, S. *et al.* (2011) Learning generative models for protein fold families. *Proteins* 79, 1061–1078

61. Lapedes, A. *et al.* (2012) Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv.org* q-bio. QM. http://arxiv.org/abs/1207.2484v1

62. Marks, D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766–20

63. Feinauer, C. and Weigt, M. (2017) Context-aware prediction of pathogenicity of missense mutations involved in human disease. *bioRxiv* http://dx.doi.org/10.1101/103051

64. Kinjo, A.R. (2017) Monte Carlo simulation of a statistical mechanical model of multiple protein sequence alignment. *Biophys. Physicobiol.* 14, 99–110

65. Hopf, T.A. *et al.* (2017) Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135

66. Sundaram, L. *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 508, 469

67. Gray, V.E. *et al.* (2018) Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Systems* 6, 116–124.e3

68. Landrum, M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067

69. Radusky, L. *et al.* (2018) VarQ: a tool for the structural and functional analysis of human protein variants. *Front. Genet.* 9, 620

70. Swett, R.J. *et al.* (2013) Hypothesis driven single nucleotide polymorphism search (HyDn-SNP-S). *DNA Repair (Amst.)* 12, 733–740

71. Beck, P. *et al.* (2012) Covalent and non-covalent reversible proteasome inhibition. *Biol. Chem.* 393, 1101–1120

72. Boomsma, W. *et al.* (2016) Bioinformatics analysis identifies several intrinsically disordered human E3 ubiquitin-protein ligases. *PeerJ* 4, e1725–18

73. Samant, R.S. *et al.* (2018) Distinct proteostasis circuits cooperate in nuclear and cytoplasmic protein quality control. *Nature* 563, 407–411

74. Kirkegaard, T. *et al.* (2010) Hsp70 stabilizes lysosomes and reverts Niemann-Pick disease-associated lysosomal pathology. *Nature* 463, 549–553

75. Kirkegaard, T. *et al.* (2016) Heat shock protein-based therapy as a potential candidate for treating the sphingolipidoses. *Sci. Transl. Med.* 8, 355ra118–355ra118

76. Pereira, D.M. *et al.* (2018) Tuning protein folding in lysosomal storage diseases: the chemistry behind pharmacological chaperones. *Chem. Sci.* 9, 1740–1752

77. Joerger, A.C. and Fersht, A.R. (2016) The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches. *Annu. Rev. Biochem.* 85, 375–404

78. Van Goor, F. *et al.* (2011) Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proc. Natl. Acad. Sci. U. S. A.* 108, 18843–18848

79. Wright, C.F. *et al.* (2018) Assessing the pathogenicity, penetrance and expressivity of putative disease-causing variants in a population setting. *bioRxiv* http://dx.doi.org/10.1101/407981

80. Jordan, D.M. and Do, R. (2018) Using full genomic information to predict disease: breaking down the barriers between complex and Mendelian diseases. *Annu. Rev. Genomics Hum. Genet.* 19, 289–301

81. Glusman, G. *et al.* (2017) Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. *Genome Med.* 9, 113

82. Fersht, A.R. *et al.* (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* 224, 771–782

83. Allen, M. *et al.* (2018) Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ Preprints* http://dx.doi.org/10.7287/peerj.preprints.27137v1