Deep mutational scanning reveals a tight correlation between protein degradation and toxicity of thousands of non-native aspartoacylase protein variants

Martin Grønbæk-Thygesen[1], Vasileios Voutsinos[1], Kristoffer E. Johansson[1], Thea K. Schulze[1], Matteo Cagiada[1], Line Pedersen[1], Lene Clausen[1], Snehal Nariya[2], Rachel L. Powell[2], Amelie Stein[3], Douglas M. Fowler[2,4,#], Kresten Lindorff-Larsen[1,#], Rasmus Hartmann-Petersen[1,#]

Running title: Multiplexed assessment of ASPA variant abundance and toxicity

Keywords: MAVE, DMS, protein folding, protein stability, protein degradation, protein quality control, proteasome, ubiquitin, Canavan's disease, neurodegeneration, chaperone

1: Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark.
2: Department of Genome Sciences, University of Washington, Seattle, WA, USA.
3: Department of Biology, University of Copenhagen, Copenhagen, Denmark.
4: Department of Bioengineering, University of Washington, Seattle, WA, USA.

#: Correspondence to: D.M.F. (dfowler@uw.edu), K.L.-L. (lindorff@bio.ku.dk), R.H.-P. (rhpetersen@bio.ku.dk).

Abstract

When the structural stability of a protein is compromised, the protein may form non-native interactions with other cell proteins and thus becomes a hazard to the cell. To mitigate this danger, destabilized proteins are targeted by the cellular protein quality control (PQC) network, which either corrects the folding defect or targets the protein for degradation. However, the details of how the protein folding and degradation systems collaborate to combat potentially toxic non-native proteins are unknown. To address this issue, we performed systematic studies on destabilized variants of the cytosolic aspartoacylase, ASPA, where loss-of-function variants are linked to Canavan's disease, an autosomal recessive and lethal neurological disorder, characterized by the spongy degeneration of the white matter in the brain. Using Variant Abundance by Massively Parallel sequencing (VAMP-seq), we determined the abundance of 6152 out of the 6260 (~98%) possible single-site missense and nonsense ASPA variants in cultured human cells. The majority of the low abundance ASPA variants are degraded through the ubiquitin-proteasome system (UPS) and become toxic upon prolonged expression. Variant cellular abundance data correlates with predicted thermodynamic stability, evolutionary conservation, and separates most known disease-linked variants from benign variants. Systematic mapping of degradation signals (degrons) shows that inherent primary degrons in ASPA are located in buried regions, and reveals that the wild-type ASPA C-terminal region functions as a degron. Collectively, our data can be used to interpret Canavan's disease variants and also offer mechanistic insight into how *ASPA* missense variants are targeted by the PQC system. These are essential steps towards future implementation of precision medicine for Canavan's disease.

Introduction

The ubiquitin-proteasome system (UPS) is responsible for the majority of intracellular protein degradation, and thus plays a vital role in maintaining protein homeostasis [1–3]. An important group of UPS substrates include proteins that are thermodynamically destabilized or otherwise unable to attain their native conformation. When the cellular protein quality control (PQC) system detects such non-native proteins, chaperones, co-chaperones and specific E3 ubiquitin-protein ligases catalyze either their refolding or degradation via the UPS [3,4]. Along with physical conditions such as temperature [5], the folding and stability of a protein is determined by its amino acid sequence [6]. Accordingly, mutations that cause alterations in the amino acid sequence may affect the folding and thermodynamic stability of the protein. However, depending on the nature of the substitution and on its position in the protein structure, the effect of single amino acid substitutions may vary greatly, from increasing stability to a partial or complete ablation of the structure, which in turn will lead to rapid degradation of the protein. Though estimates vary depending on which approach is used [7], about half of disease-causing missense variants are thought to lead to protein destabilization and degradation [7–14]. Despite recent progress in computational biology, predicting the effects of missense variants remains a significant challenge, which in turn reduces our ability to accurately identify pathogenic gene variants [15–17].

Aspartoacylase (ASPA) (EC 3.5.1.15; UniProt entry P45381) is a 313 residue enzyme, mainly expressed in the oligodendrocytes of the brain [18–20]. Here, it facilitates the hydrolysis of N-acetyl-aspartate (NAA), one of the most abundant amino acid-derived metabolites in the brain and a marker of brain health [21], into aspartate and acetate [22,23]. Structurally, the protein consists of a single domain with a channel leading to the active site [23]. Various interactions around the substrate binding cavity ensure high substrate specificity. Additionally, coordinated by the conserved residues H21, E24, and H116, ASPA binds a Zn^{2+} ion [23,24]. Although the monomeric enzyme is active [25], several studies have demonstrated that ASPA forms homodimers [19,26,27] through a large interaction surface [23,24].

Insufficient ASPA activity caused by germline *ASPA* variants is linked to Canavan´s disease (CD) (OMIM: 271900), a recessive, neurodegenerative leukodystrophy, in which oligodendrocytes fail to properly myelinate neuroaxons [28]. The mechanism of pathogenicity remains somewhat obscure, with two suggested hypotheses that are not mutually exclusive: the osmotic-hydrostatic hypothesis suggests that abnormal NAA build-up in CD patients, and the resulting osmotic consequences, cause the disease [28,29]. According to the acetyl-lipid myelin hypothesis, insufficient acetate production from NAA, which is normally incorporated into the myelin sheath, is the underlying cause [18,29,30].

Clinically, CD patients suffer from poor muscle control, reduced cognitive capabilities and other severe conditions. The symptoms appear within the first 3–6 months of life, and worsen over time eventually leading to an early death [31–33]. Various attempts at curing CD or ameliorating the symptoms [34–38] have been reported, with gene therapy being one of the more promising [39–41]. Two clinical trials aiming at curing Canavan´s disease with gene therapy are currently ongoing (ClinicalTrials.gov Identifier: NCT04998396 and NCT04833907). Due to the progressive nature of the disease, such an intervention would likely need to be performed early [41–43], thus making it essential to rapidly assess whether novel *ASPA* variants are pathogenic.

In a previous study, we have shown that the disease-linked ASPA C152W variant is targeted for PQC-linked degradation in both yeast and human cells [44]. To further investigate the PQC degradation of ASPA in relation to Canavan´s disease, we here generated a site saturated library of *ASPA* variants and analyzed it using the variant abundance by a massively parallel sequencing (VAMP-seq) technique [45]. The resulting variant effect map comprises 6152 out of the 6260 (19 substitutions/residue*313 residues + 312 early stop codons + 1 wild-type) corresponding to ~98%

of all possible single-site missense and nonsense *ASPA* variants, and correlates with thermodynamic stability predictions and evolutionary conservation. Many of the low abundance ASPA variants are toxic to the cells, indicating a tight connection between structural destabilization, PQC-linked degradation and toxicity. Our data emphasize the importance of low ASPA abundance as a major mechanism of Canavan´s disease and reveals an intimate link between reduced thermodynamic stability, degradation and toxicity.

Results

A massively parallel assay for ASPA protein abundance

We have previously shown that the disease-linked C152W ASPA variant is subject to chaperone-dependent proteasomal degradation [44]. To test whether this is a common trait for disease-linked ASPA variants, we applied variant abundance by massively parallel sequencing (VAMP-seq) [45] to a site saturated and barcoded cDNA library of *ASPA* variants. In our approach, the ASPA library consists of ASPA variants fused to GFP after site-specific recombination at a "landing pad" locus in human HEK293T cells (Fig. 1A). Since the plasmid does not contain a promoter, non-integrated plasmids are not expressed, while correct Bxb1-catalyzed site-specific integration at the landing pad locus leads to single-copy expression of GFP-fused ASPA. To correct for cell-to-cell variations in expression, the integrated plasmid also produced mCherry from an internal ribosomal entry site (IRES) downstream of ASPA. Fluorescence-activated cell sorting (FACS) is used to separate cells into distinct bins based on the GFP:mCherry ratio, and then high throughput DNA sequencing is used to quantify the frequency of every variant in each bin by sequencing the barcodes (Fig. 1A). Since integration of the plasmid at the landing pad will block expression of BFP and iCasp9 [46], non-recombinant cells can be identified based on expression of BFP and depleted from the culture by adding AP1903 (Rimiducid), which specifically induces apoptosis of iCasp9 positive cells (Fig. 1A).

To test the feasibility of the assay, we initially compared wild-type (WT) ASPA and the C152W disease-linked variant, which was previously shown to be rapidly degraded via the proteasome and therefore of low abundance [44]. Indeed, fluorescence microscopy revealed dramatically reduced abundance of the C152W variant (Fig. 1B). This was also evident by western blotting and was independent of whether GFP was fused to the N-terminus or C-terminus of ASPA (Fig. 1C).

However, to also allow for analyses of nonsense variants, we proceeded with the GFP fused to the N-terminus of ASPA.

By flow cytometry, the mCherry levels appeared similar for WT and C152W, while the GFP level of WT was roughly 10-fold greater than that of C152W (Fig. 1DE). Finally, since we were unable to detect endogenous ASPA in the HEK293T cells (Fig. S1), the GFP-ASPA protein abundance is likely independent of potential heterodimer formation with endogenous ASPA.

Comprehensive mapping of ASPA variant protein abundance

We generated a site saturated library of *ASPA* missense and nonsense variants and inserted it in frame with GFP (Fig. 1A). An oligo containing 18 random nucleotides was also inserted in the plasmid to serve as a barcode for the subsequent analyses (Fig. 1A). The resulting plasmid library was then subjected to long-read PacBio sequencing. This allowed us to match 134,176 unique barcodes with individual *ASPA* variants corresponding to each variant being represented on average by ~21 different barcodes. Thus, for all subsequent experiments, variants could be identified by short-read Illumina sequencing of the barcodes.

The barcoded ASPA library was transfected into the HEK293T cell line and cells in which no recombination at the landing pad locus had occurred were eliminated with AP1903. By flow cytometry, GFP:mCherry levels in the library spanned more than an order of magnitude and covered the range between the WT and C152W controls (Fig. 1DE). FACS was used to separate the cells into four equally populated bins based on the GFP:mCherry levels (Fig. 1F). Then, Illumina sequencing of the barcodes allowed us to quantify the frequency with which each variant is found in each of the four bins (Fig. 1AF) and calculate an abundance score ranging from 1 (WT-like abundance) to 0 (strongly reduced abundance). The average Pearson correlations between replicate experiments was 0.99 (Fig. S2). However, we note that for the low score variants the correlations between replicates were not as strong, indicating a poor resolution for the low abundance variants. The final scores and standard deviations were determined based on 11 replicates, and revealed the relative abundance of 5843 out of 5947 (98%) missense variants and 308 out of 312 (99%) nonsense variants (Fig. 2A). The abundance scores were bimodally distributed, with a WT-like peak of stable variants centered on the synonymous (silent) substitutions and a peak of low-abundance variants centered on the nonsense variants (Fig. 2B). The abundance of 18 different ASPA variants, determined individually by flow cytometry in low throughput, were consistent with the high throughput map (Fig. 2C). We observe a poorer resolution of the lowest abundance variants, which appears—at least in part—to be connected to the reduced fitness observed for many of these variants (see below and discussion). The median abundance score per position (shown in Fig. 2A) explained 55% of the total variance of the variant abundance scores (equivalent to a Pearson correlation of 0.77). Thus, the tolerance to amino acid substitutions appeared more dependent on position than the nature of the target amino acid. However, as expected, substitutions to proline appear detrimental at most positions (Fig. 2A). The map revealed that most regions of ASPA are sensitive to substitutions, although a particular loop stretching from position 70 to 110 appeared more tolerant (Fig. 2A), while many variants in the disordered regions (as predicted by low AlphaFold pLDDT scores) near the N- and C-termini displayed an increased abundance. In the loop from position 159 to 166, substitutions to hydrophobic residues reduce ASPA abundance (Fig. 2A).

When mapping the median abundance scores at each position onto the ASPA structure, some surface regions appeared sensitive to substitutions (Fig. 2D). However, most regions buried in the core of the ASPA structure were highly sensitive to substitutions (Fig. 2D), including the residues coordinating the Zn^{2+} ion at the active site (Fig. S3A). For the exposed β-strand at position 150-160, the substitutions alternated between destabilizing and stabilizing, corresponding to residues pointing inwards and outwards, respectively (Fig. S3B). Accordingly, many of the low abundance positions are buried in the structure and display a high weighted contact number (WCN) [47] (Fig. S4). However, some low abundance exposed positions were evident, corresponding to the sensitive regions on the ASPA surface (Fig. 2D).

The abundance of ASPA variants correlates with predicted thermodynamic folding stability

Previous reports on other proteins, have suggested that variant protein abundance correlates with the experimental or predicted thermodynamic stability of the folded protein [45,48,49]. To probe this relationship for ASPA, we next employed structure-based energy calculations to predict the effects of missense variants on the thermodynamic (structural) stability of ASPA. Using the published crystal structure of the ASPA homodimer (PDB: 2O53) [24], and introducing all possible single amino acid substitutions, we applied the Rosetta energy function [50] to estimate the change (Δ) in thermodynamic folding stability (ΔG) compared to wild-type ASPA (ΔΔG). In total the data comprise 5719 variants (19 possible amino acid substitutions per position * 301 positions resolved in the structure). The resulting ΔΔG values report on the predicted change in thermodynamic stability of the ASPA dimer, such that variants with ΔΔGs close to zero represent a WT-like stability, while variants with large positive ΔΔGs should be less stable than WT ASPA, and have a higher proportion of fully or partially unfolded structures that are targeted for degradation. A comparison of the Rosetta predictions with the abundance scores represented as heat maps is included in the supplemental information (Fig. S5AB). Overall, the thermodynamic stability predictions correlated with the experimental abundance scores (Spearman's $\rho = -0.47$), which were further strengthened when comparing the median values per residue (Spearman's $\rho = -0.55$) (Fig. 3A). However, some variants were either predicted to be unstable (high ΔΔG) but observed at high abundance, or predicted as stable (low ΔΔG) but observed at low abundance (Fig. 3A). Hence, the thermodynamic stability predictions capture some, but not all, of the observed effects. For instance, surface exposed sensitive regions or substitutions that introduce degrons, will not be captured. Comparisons of the abundance scores with the Rosetta predictions based on the ASPA monomer (Fig. S5AB) and the difference (Δ(ΔΔG) between the Rosetta predictions for the monomer and dimer (Fig. S5AB), did not reveal any abundance effects which could be attributed directly to dimer formation.

The abundance of ASPA variants correlates with evolutionary conservation

In folded proteins, residues critical for function *e.g.* those in the active site and/or for maintaining the native structure, are typically highly conserved across different species. Accordingly, sequence conservation across ASPA orthologues should predict the mutational tolerance of the protein at the residue level. To test this, we first generated a multiple sequence alignment of 757 different ASPA homologues and then applied the GEMME [51] model that takes into account both residue conservation and the non-trivial pair couplings that occur as a consequence of amino acid co-variation. The resulting evolutionary distance scores report on the likelihood of a given substitution, where a score close to zero indicates a neutral variation with no effect on the structure and/or function of the protein. Conversely, substitutions with large negative GEMME scores are predicted as unfavorable. Again, we observed a correlation (Spearman's $\rho = 0.45$) between the experimental abundance scores and the predictions (Fig. 3B). As these sequence-based predictions do not discriminate between residues that are conserved for function or structure, many of the outliers in our correlations may simply be residues that are important for function but do not contribute to thermodynamic stability of the native fold. A comparison of the GEMME predictions with the abundance scores, and Rosetta stability predictions are represented as heat maps is included in the supplemental information (Fig. S6AB).

Most destabilized ASPA variants are heat-labile PQC and proteasome targets

Next, we proceeded to explore the cellular and physical mechanisms causing the low abundance. To this end, cells transfected with the ASPA library were subjected to a range of physical and chemical perturbations while following the distribution of the variants by flow cytometry. The flow cytometry profiles of the WT, C152W and the variant library in unperturbed cells are shown for comparison (Fig. 4A). First, we noted that the flow cytometry profiles for cells incubated at 29, 37 or 39.5 °C differed widely. Thus, at 39.5 °C the unstable peak became more pronounced (Fig. 4B), which suggests that at this temperature, variants with low or intermediate abundance at 37 °C are further destabilized. At 29 °C, however, the low abundance peak was reduced to a small shoulder indicating that most variants were stabilized (Fig. 4C).

Treating the cells with an inhibitor of the E1 ubiquitin-activating enzyme (MLN7243) led to an increased abundance (Fig. 4D). Accordingly, an increased abundance was also evident when the proteasome was blocked with bortezomib (BZ) (Fig. 4E). Conversely, there was little effect of treating the cells with autophagy-inhibitor chloroquine (CQ) (Fig. 4F). This indicates that most of the low abundant ASPA variants are targeted by the ubiquitin-proteasome system, while autophagic clearance of ASPA variants is insignificant.

Since HSP70-type molecular chaperones have been shown to play an important role in PQC-linked degradation of destabilized proteins, including ASPA C152W [44], we tested the effect of the HSP70 inhibitor, YM01. Similar to the situation with bortezomib, the HSP70 inhibitor shifted the unstable peak towards a higher GFP:mCherry ratio (Fig. 4G), indicating that HSP70 plays a role in the degradation of many ASPA variants.

Finally, since ASPA folding and stability could potentially be affected by substrate binding, we also analyzed the library distribution in the presence of NAA. However, no effects were evident upon adding NAA to the cells (Fig. 4H), though we note that the resulting intracellular concentration of NAA is unknown.

Inherent PQC degrons in ASPA map to buried regions that are sensitive to mutation

The reigning hypothesis explaining the degradation of destabilized or misfolded proteins, states that these proteins, through local or global unfolding events, transiently expose PQC degradation signals (degrons). Degrons are recognized by E3 ubiquitin-protein ligases and/or molecular chaperones such as HSP70 [52–55], which in turn direct the protein for proteasomal degradation.

Given the effect of inhibiting the ubiquitin-proteasome system on ASPA variants, we reasoned that ASPA likely contains PQC degrons, and that mapping these degrons could shed additional light on the ASPA abundance map. To identify degrons, the ASPA sequence was divided into 24-residue tiles each overlapping by 12 residues (Fig. 5A). Similar to full-length ASPA, a library of the tiles was fused to the C-terminus of GFP and expressed from the landing pad in the HEK293T cells. The cells were flow sorted and sequencing across the tiles revealed the frequency of each ASPA tile in the four different bins (Fig. 5B). Ultimately, this allowed us to calculate a tile stability index (TSI) covering the ASPA sequence (Fig. 5C). Indeed, multiple tiles display a low TSI and thus had reduced GFP:mCherry levels, suggesting that these tiles harbor degrons. When comparing with the ASPA structure, the low abundance tiles generally appeared buried in the structure (Fig. S7). Accordingly, regions with low TSI also partly overlapped with regions that display a high average weighted contact number (Fig. 5C). Comparing the mapped TSIs with PQC degron predictions made with the quality control degron predictor (QCDPred) [56,57], revealed that regions displaying a low TSI also displayed a high QCDPred degron probability, while regions with a low degron probability appeared stable (Fig. 5D). This indicates that the sequence features of the ASPA degrons are similar to those reported for PQC degrons in general, *i.e*. enriched in hydrophobic residues and depleted for acidic residues [56–59]. Finally, the C-terminal tile displayed degron properties (Fig. 5C). This may suggest that the ASPA contains a C-degron [58,59] or that the C-terminal region functions as a disordered degradation initiation site (tertiary degron) [60,61], but due to the high QCDPred score could also reflect a PQC degron. We note that both substitutions and truncations in the ASPA C-terminus increase ASPA abundance (Fig. 2A), supporting the presence of a degron at this position in wild-type ASPA.

Most disease-linked ASPA variants have reduced abundance

Next, we examined if the abundance map could distinguish known harmless (benign) and disease-linked ASPA variants. Based on the ClinVar database [62] and frequency in the population, as reported in the Genome Aggregation Database (gnomAD) [63], we first collected a curated list of disease-linked and benign ASPA variants (SupplementalFile1.xlsx). The three variants listed in ClinVar as benign/likely benign and also observed most frequently in the population, all displayed an abundance similar to wild-type ASPA (Fig. 6A). Conversely, 50 out of 61 pathogenic variants displayed an abundance score lower than 0.5 (Fig. 6A). Among the remaining pathogenic high-abundance variants, most were at catalytic sites (Fig. 6A, blue markers). Thus, these variants are likely pathogenic due to inactivating ASPA function without affecting ASPA thermodynamic stability and abundance. The so-called variants of uncertain significance (VUS), *i.e.* variants for where a clinical interpretation is currently lacking, clustered into high and low abundance groups (Fig. 6A). We suggest that those with low abundance are likely to be pathogenic.

As expected, comparing the abundance scores with the allele frequencies of the ASPA variants reported in gnomAD, revealed that the most common ASPA missense alleles are benign and display wild-type like abundance scores, while most of the low abundance variants are rare (Fig. 6B).

Then, we examined the abundance score for the clinical variants in combination with the evolutionary conservation scores generated with GEMME (Fig. 6C). The high GEMME scores for benign and highly abundant variants indicate that these substitutions occur at evolutionary tolerant sites, indicating that they are likely functional and stable proteins. For pathogenic variants, there was a lower match with evolutionary conservation. For the 11 highly abundant pathogenic variants (abundance score > 0.5), five showed a high level of evolutionary conservation (GEMME score < -3), suggesting they play a role in enzyme activity. Additionally, most low abundance, pathogenic variants are further distinguished from benign, abundant variants by the GEMME score.

Certain ASPA variants become toxic upon prolonged expression

While conducting the abundance experiments presented above, we noticed that upon prolonged incubation after inducing expression by the addition of doxycyclin, cells transfected with ASPA C152W were lost from the cultures while the small group of BFP positive cells that survived treatment with AP1903 instead became more profuse (Fig. S8A). Conversely, cells expressing wild-type ASPA did not disappear from cultures for at least up to 9 days (Fig. S8B). Since this suggests that some non-native ASPA protein species (including C152W) become toxic upon prolonged expression, we next screened the library for such toxic variants. To this end, the library was introduced into the landing pad in the HEK293T cells and non-recombinant cells were eliminated with AP1903. Then, without flow sorting, the cells were harvested after 0, 5, 7 and 9 days in culture and the surviving variants identified by sequencing of the barcodes (Fig. S8C). Thus, by following the propagation of each variant in the culture over time, we could quantify the competitive fitness of each variant and calculate a toxicity score ranging from 1 (toxicity similar to C152W) to 0 (non-toxic similar to WT). The final toxicity scores and standard deviations were determined based on four biological replicates and the average Pearson correlation between replicate experiments was

0.93 (range: 0.93-0.94) and the mean absolute error 0.10 (Fig. S9). Since the coverage is mainly limited by library synthesis, the coverage of toxicity scores was similar to the abundance score coverage with 5847 of 5947 (98%) missense variants and 307 out of 312 (98%) nonsense variants (Fig. 7A). The toxicity scores displayed a bimodal distribution with a peak overlapping with the synonymous WT non-toxic variants and smaller peak of toxic variants (Fig. S10). Comparing the toxicity scores with the abundance scores revealed that all toxic variants were low abundance variants (Fig. 7B), indicating that continued expression of some destabilized ASPA variants is toxic, resulting in a gradual depletion of such variants from the population. The abundance levels measured in low throughput of toxic variants are lower than for non-toxic variants (Fig. S11). This suggests that the toxicity scores may reflect the abundance scores, thus improving resolution of the low-abundance variants in the abundance screen (Fig. 2C). Accordingly, we observe a correlation (Fig. 7C) between variant toxicity and Rosetta ΔΔG values (Spearman's ρ = 0.53), indicating that toxic variants tend to be more thermodynamically destabilized than non-toxic variants. In particular, many toxic variants are highly destabilized in our assay. However, not all variants with large positive ΔΔG values are toxic. Likewise, the toxic variants also appeared more unfavorable in the GEMME-based evolutionary conservation analyses (Fig. 7D). Most of the nonsense variants were non-toxic (Fig. 7A and Fig. S10), while generally of low abundance (Fig. 2AB). A side-by-side comparison of the toxicity, abundance, Rosetta ΔΔG and GEMME maps is provided in the supplemental material (Fig. S12). Similar to the abundance map, the positions where most substitutions resulted in toxic ASPA variants were found in regions buried within the ASPA structure (Fig. 7E). Accordingly, we note a correlation (between the toxicity score and the weighted contact number (WCN) (Fig. S13) and a partial overlap between toxic positions and the mapped degrons (Fig. S12B). None of the benign variants were toxic, while the pathogenic variants clustered into toxic and non-toxic groups (Fig. S14). Hence, we predict that highly toxic VUS are likely to be pathogenic.

Toxic, low abundance variants trigger a stress response leading to induction of HSP70

Based on the results above, we conclude that some of the thermodynamically destabilized and low abundance ASPA variants reduce cellular fitness upon prolonged expression. To further characterize this effect, we compared the transcriptomes of cells expressing the WT (non-toxic) and C152W (toxic) variant by RNA sequencing. Principal component analysis (PCA) revealed that the three independent WT samples clustered together, whereas the C152W samples were more spread out, indicating a larger variation between the toxic samples (Fig. S15A). Among the differentially expressed genes we observed the small heat-shock protein HSPB8 and the HSP70-type chaperone HSPA1B as significantly upregulated in cells expressing C152W (Fig. S15B). Among the Gene Ontology (GO) terms that were significantly enriched, we noted several related to cell stress and apoptosis (SupplementalFile4.xlsx), indicating that the toxicity is linked to a stress response caused by the degradation of thermodynamically unstable ASPA variants. Therefore, we tested if the expression of selected ASPA variants would activate the stress response pathway by measuring the induction of the stress responsive HSPA1B by qPCR. Indeed, we find a correlation between abundance, toxicity and HSPA1B mRNA levels (Fig. S15CDE), suggesting that the reduced fitness of certain low abundant ASPA variants is connected with activation of the stress response pathway, which in turn inhibits cell growth.

Discussion

In the present work we probed the intimate relationship between missense protein variants, thermodynamic folding stability, degradation and toxicity. By draining the cell for resources and through formation of non-specific interactions, the expression of non-native proteins has for long been recognized to be toxic to cells [64–66]. However, in most cases, studies on this have been limited to the expression of a single toxic protein such as Huntingtin or α-synuclein, etc. [67,68], genetically linked to dominant diseases. Since Canavan's disease is a recessive disorder [32], the toxicity of certain low abundance ASPA variants is unlikely to contribute to the disease, but rather a consequence of the variants being overexpressed which provides us with a glimpse of how the PQC network operates. Some toxic and misfolded protein species form aggregates, and although ASPA C152W, when expressed in yeast cells, localizes to large cytosolic inclusions [44], we never observed any aggregates in HEK293T cells. However, we cannot exclude that smaller inclusions are formed, which may influence turnover and toxicity. RNA sequencing revealed that the toxic C152W variant led to a differential expression of genes involved in stress response and apoptosis, including an upregulation of the of stress-responsive HSP70-type chaperone HSPA1B. As the HSPA1B induction correlated with a reduced abundance and increased toxicity, this suggests that the toxicity is caused by the presence of destabilized ASPA variants that are prone to misfold and therefore rapidly turned over.

Most likely, the toxicity of low abundant variants is not unique for ASPA. In a parallel study, we have analyzed the abundance of a saturated library of variants in the protein Parkin [69]. For that protein, we did not observe any toxic effects of low abundance variants. It is possible that the low abundance ASPA variants are expressed at a higher level than the Parkin variants, and therefore potentially burdening the PQC system more severely. However, since Parkin is a modular protein, composed of multiple smaller domains, while ASPA is a large single domain protein, it is also possible that ASPA unfolding events will affect the protein globally and thus be more dramatic than a local unfolding event localized to a single domain in Parkin. In agreement with this, we note that while most ASPA nonsense variants are of low abundance, the toxic nonsense variants primarily cluster towards the C-terminal region, indicating that the toxicity primarily occurs when the bulk of ASPA has been produced.

Knowing that some variants can be toxic to the cells will be important for other VAMP-seq screens and other implementations of multiplex assessment of variant effects (MAVE) technologies. When considering how toxicity could impact the abundance experiment, we note the high density of variants around abundance score zero (Fig. 2B) and the elevated uncertainty in this region (Fig. S2) may in part be a consequence of the toxicity because toxic variants are poorly represented in terms of the number of cells. Thus, the average Spearman replica correlation of 951 non-toxic variants (toxicity score < 0.4) with abundance score < 0.2 is 0.77, but only 0.09 for 1529 toxic (toxicity score > 0.6) variants also with abundance score < 0.2. This substantial difference in replica correlations supports that the toxicity causes poor resolution in the low abundance region also observed for the low throughput validation experiments (Fig. 2C). Thus, although the toxicity should not affect the abundance scores directly, it may result in reduced resolution among low-abundance variants because toxicity only affects low-abundance variants and because the FACS gates are set to include the same number of cells in each bin.

The observed correlation between protein abundance and thermodynamic folding stability suggests that most low-abundance ASPA variants are thermodynamically destabilized in their structure. In turn, this will cause such variants to more frequently populate fully or partially unfolded states where regions that are buried in the native conformation become exposed. Recent structural studies on disease-linked variants in dihydrofolate reductase showed that structural destabilization led to transient exposure of a PQC degron [54]. We show that ASPA also contains multiple regions that lead to degradation when artificially exposed by grafting them as peptides onto GFP. Since most of these regions are buried in the native conformation and have sequence properties leading to high scores with QCDPred, we suggest that these fragments work as PQC degrons. Presumably, at least some of these degrons are involved in targeting non-native ASPA variants for degradation via the UPS.

Although we find that HSP70 contributes to the ubiquitin-dependent proteasomal turnover of many ASPA variants, we do not presently know the identity of the UPS components, including E3s, which mediate the degradation. Even though one candidate is the E3 ligase CHIP, which is known to target certain HSP70 clients for proteasomal degradation [70–72], data from yeast cells indicate a high level of redundancy between the E3s linked to the degradation of PQC substrates [73–75]. Accordingly, matching non-native proteins with their corresponding PQC E3s is not straightforward, but highly important since inhibiting such E3s should lead to increased levels of destabilized variants. In turn, this could potentially broadly alleviate genetic disorders where the variant proteins, albeit structurally destabilized, are still functional. Indeed, many disease-linked protein missense variants that are targeted for PQC-linked degradation are still, at least partially, functional [54,76,77], indicating that the PQC system is tightly tuned to root out non-native protein species. An alternative approach would be to develop small molecule stabilizers/correctors that through binding to the native conformation could block PQC-linked degradation and reactivate certain pathogenic variants. Indeed such drugs have been successfully developed and implemented for cystic fibrosis [78].

Gene therapy is currently one of the more promising attempts at curing Canavan's disease [39–41,79]. However, as the disease is highly progressive, such an intervention would most likely need to be performed early [41–43]. Consequently, it is essential with better tools to assess whether VUS and novel *ASPA* variants are pathogenic. To that end, MAVE assays, like those presented here, offer comprehensive genotype-phenotype information [80]. The presented data therefore provide an essential step towards a future implementation of gene therapy or precision medicine approaches for Canavan's disease.

Materials and methods

Plasmids and library creation

The wild-type *ASPA* cDNA and selected variants studied in low throughput were generated by Genscript. The library cloning and barcoding described below are essentially as previously described [69]. The *ASPA* site-saturation mutagenesis library was purchased from Twist Biosciences and resuspended in 50 µL nuclease free water to a final concentration of 100 ng/µL. Then, two independent 50 µL reactions with 1 µg of backbone plasmid were digested at 37 °C for 1 hour with MluI-HF and EcoRI-HF (New England Biolabs). After heat-inactivation (65 °C, 20 min), the products were purified following manufacturer's protocols by resolving on a 1 % agarose gel with SYBR Safe (ThermoFisher Scientific), followed by a gel extraction (Qiagen) and cleanup (Zymo Clean and Concentrate) of the 5.3 kb band. The product was assembled with the library oligonucleotide (diluted ten-fold) in a Gibson reaction (insert:backbone molar ratio of 2:1) at 50 °C for 1 hour. The assembly products were then cleaned and eluted in 6 µL water (Zymo Clean and Concentrate). Then, 1 µL of Gibson assembly product was incubated with 25 µL *E. coli* NEB-10β cells for 30 min on wet ice, prior to electroporation (2 kV, 6 millisec). The cells were resuspended in 975 µL SOC (Sigma) immediately after electroporation and incubated at 37 °C for 1 hour with gentle agitation. Subsequently, 1 mL culture was used to inoculate 99 mL LB media containing 100 µg/mL ampicillin and grown overnight in a 37 °C. In addition, and prior to the overnight growth, to estimate library coverage by colony count, 100 µL, 10 µL, and 1 µL samples were collected and spread on LB agar plates containing 100 µg/mL ampicillin. After the overnight growth at 37 °C, the cells were harvested by centrifugation (30 min, 4300 g) and plasmid purified by midi-prep (Millipore Sigma).

Library barcoding

For barcoding of individual variants, 1 µg of the library plasmid was digested at 37 °C for 5 hours with NdeI-HF and SacI-HF (New England Biolabs). Then, 1 µL rSAP was added for 30 min. at 37 °C, followed by a 20 min heat-inactivation at 65 °C. The digested library was purified by 1% agarose gel electrophoresis, followed by gel extraction (Qiagen). Next the library vectors were further purified and eluted in 10 µL water using the Zymo Clean and Concentrate kit.

The barcoding oligonucleotides, which contained 18 degenerate nucleotides, (IDT) were resuspended in water to a concentration of 10 µM. In order to anneal the barcode oligo, 1 µL of oligo was added to a mix of 1 µL 10 µM MAC356 primer, 4 µL CutSmart buffer, and 34 µL water. This reaction was incubated at 98 °C for 3 minutes, and then ramped down at -0.1°C/s to 25 °C. To fill in the barcode oligo, 1.35 µL of 1 mM dNTPs and 0.8 µL Klenow exo-polymerase (New England Biolabs) were added and incubated at 25 °C for 15 min, 70 °C for 20 min, then ramped down to 37°C at -0.1°C/s. Once the temperature was 37 °C, the product was digested for 1 hour with 1 µL each of NdeI-HF, SacI-HF, and CutSmart buffer. Lastly, the digested product was run on a 2% agarose gel with 1x SYBR Safe (Thermo Fisher Scientific) extracted by a gel extraction kit (Qiagen) and purified further (Zymo Clean and Concentrate), followed by elution in 30 µL water.

To ligate the barcoded oligonucleotides, a ratio of 7:1 (oligo:library) was used overnight at 16 °C with T4 DNA ligase (New England Biolabs). The products were purified and eluted in 6 µL water (Zymo Clean and Concentrate). Using the same procedure as above, *E. coli* NEB-10β cells were electroporated with 1 µL of ligation product. To bottleneck the library-barcode ligation product, electroporation recovery volumes of 500 µL, 250 µL, 125 µL, and 40 µL were separately used to inoculate 50 mL LB media containing 100 µg/mL ampicillin. Then, 100 µL, 10 µL, and 1 µL samples were spread on LB agar plates containing 100 µg/mL ampicillin, for each of the 50 mL cultures, to estimate library coverage. After growth overnight at 37 °C, the library coverage was estimated by counting colony-forming units (CFU). After overnight growth at 37 °C, each of the 50 mL cultures were centrifuged (30 min, 4300 g) and plasmid purified by midi prep (Millipore Sigma). The bottlenecked library displayed an estimated 16.9 fold barcode/variant coverage.

Subassembly of barcode-variant map by PacBio sequencing

Using the enzymes XmaI and NdeI-HF (New England Biolabs), 5 μg of barcoded library was digested in CutSmart buffer at 37 °C for 5 hours. This was followed by heat inactivation at 65 °C for 20 min. The digested products were then purified with AMPure PB beads (Pacific Biosciences, 100-265-900). Library preparation and DNA sequencing were performed by the University of Washington PacBio Sequencing Services. Throughout, DNA quantity was tested with fluorometry on a DS-11 FX instrument (DeNovix) with the Qubit dsDNA HS Assay Kit (Thermo Fisher). Sizes were analyzed on a 2100 Bioanalyzer (Agilent Technologies) with the High Sensitivity DNA Kit. SMRTbell sequencing libraries were generated using the protocol ‘Procedure & Checklist - Preparing SMRTbell libraries using PacBio Barcoded Universal Primers for Multiplexing Amplicons’ and the SMRTbell Express Template Prep Kit 2.0 (p/n 100-938-900). The SMRTbell libraries were size-selected to remove backbone fragments using the SageELF (SageScience). The libraries were bound with Sequencing Primer v4 and Sequel II Polymerase v2.1 and sequenced on one SMRT Cell 8M using Sequencing Plate v2.0, diffusion loading, pre-extension for 1 hour, and a movie time of 30 hours. Calculation of CCS consensus was performed using SMRT Link version 9.0 set at the default settings. Only reads passing an estimated quality filter of ≥Q20 were selected as “HiFi” reads.

Finally, the barcoded libraries were pooled by normalizing mass to the number of constructs contained in each pool. Then the library was bound with Sequencing Primer v4 and Sequel II Polymerase v2.0. Sequencing was performed using SMRT Cells 8M using Sequencing Plate v2.0, diffusion loading, a 90 min pre-extension, and a 30 hour movie time. Further data were collected after SMRTbell Cleanup Kit v2 treatment to remove imperfect templates, with Sequel Polymerase v2.2. Adaptive loading with a target of 0.85, and a 1.3 hour pre-extension time was used. CCS consensus and demultiplexing were performed using SMRT Link version 10.2 set at default. Reads that passed an estimated quality filter of ≥Q20 were selected for mapping barcodes to variants. PacBio reads were filtered for reads with less than ten CSS passes using samtools version 1.16 [81] and aligned to the barcode-GFP-ASPA construct using BWA version 0.7.17 [82]. The barcode and ASPA sequences were extracted using cutadapt version 3.2 [83], see pacbio/pacbio_align.sh available on GitHub. Reads containing ten or more DNA substitutions or any indels were filtered out. In cases where multiple ASPA variants mapped to the same barcode, the variant with most read counts was used. This resulted in a barcode map of 134,176 unique barcodes, see pacbio/barcode_map.r on GitHub. Of these, 5,970 are wild-type, 6,122 are synonymous wild-type, and 119,301 are single amino-acid variants including 5% nonsense variants. More than 98% of all possible single amino acid substitutions (incl. stop) are covered by this library and 301 of 313 positions are fully covered. Only position 1, 188, 189, 242 and 243 were missing more than 2 substitutions and the majority of these were not synthesized in the library. Code is available at https://github.com/KULL-Centre/_2023_Groenbaek-Thygesen_ASPA_MAVE and sequencing reads at https://doi.org/10.17894/ucph.3e05fe3a-4d7e-4d70-9056-18ed999e7e1e.

Cell propagation, transfection and recombination

Experiments were performed using HEK293T landing pad cell line TetBxb1BFPiCasp9 Clone 12, which was characterized previously [46]. The cells were maintained in Dulbecco’s Modified Eagle´s Medium (DMEM) (Sigma-Aldrich) supplemented with 10 % fetal bovine serum (Sigma), 64.43 mM Penicillin G (AppliChem), 27.45 mM Streptomycin sulfate (AppliChem), and 2 mM glutamine (Sigma), with 2 µg/mL doxycycline (Dox) (Sigma-Aldrich), and split at around 80-90 %

confluency. For recombination of libraries, 3.5 million cells were seeded out into a 10 cm plate with 10 mL DMEM without doxycycline. The next day, the cells were transfected as follows: In one tube, 7.1 µg library DNA and 0.48 µg pNLS-bxb1-recombinase was mixed with OptiMEM (Gibco) in a total volume of 710 µL. In another tube, 28.5 µL Fugene HD (Promega) was added to 685 µL OptiMEM and mixed gently. The two solutions were mixed gently by pipetting and incubated for 15 minutes before being added to the cells in 10 cm plates. Approximately 48 hours later, 2 µg/ml doxycycline and 10 nM AP1903 (MedChemExpress) was added. The library was grown for 5 days in doxycycline before FACS profiling/sorting. A minimum of 10^6 cells, corresponding to approximately 150 fold coverage of the total number of variants, were maintained in the population at all times following library recombination.

SDS-PAGE and western blotting

Cells were washed in PBS and then harvested in SDS sample buffer (3% SDS, 93 mM Tris/HCl pH 6.8, 18 % glycerol, 0.02% Bromophenol blue, 2.5% (v/v) 2-mercaptoethanol) and boiled at 100 °C for 2 minutes. The samples were resolved on 12.5% acrylamide separation gels with a 3 % (w/v) stacking gel using a constant voltage of 125 V for approximately 1 hour in running buffer (50 mM Tris, 0.4 M glycine, 0.1% SDS). Next, the proteins were transferred onto a nitrocellulose membrane (pore size 0.2 µm) (Advantec), in-between filter papers (Frisenette) soaked in transfer buffer (50 mM Tris-base, 100 mM glycine, 0.01 % SDS, 20 % (v/v) ethanol), at 100 mAmp/gel for 1.5 hours. Transferred proteins were stained in Ponceau S (0.1% Ponceau S (Sigma-Aldrich), 5% (v/v) acetic acid), Excess Ponceau was washed away with PBS (0.137 M NaCl, 2.68 mM KCl, 6.46 mM Na_2HPO_4, 1.47 mM KH_2PO_4, pH 7.4), before areas of interest were excised out of the membrane. The excised membrane pieces were incubated in blotto buffer (5% fat-free milk power in PBS) for at least 30 minutes, and incubated with primary antibody overnight. Following 3 rounds of 10 minute incubations in wash buffer (50 mM Tris/HCl pH 7.4, 150 mM NaCl, 0.01% (v/v) Tween-20) the blots were incubated in horse radish peroxidase (HRP)-conjugated secondary antibodies for 1 hours. After, 3 rounds of 10 minute incubations in wash buffer, the blots were developed using chemiluminescence (GE Healthcare, 1059243, 1059250) on a BioRad ChemiDoc MP Imaging System imager (BioRad, 12003154). The primary antibodies used and their sources were: rabbit anti-aspartoacylase (Thermo Scientific, PA5-29180) (diluted 1:1000), mouse anti-β-actin (Sigma-Aldrich, A5441) (diluted 1:25.000), rat anti-GFP (Chromotek, 3H9) (diluted 1:1000), mouse anti-RFP (Chromotek, 6G6) (diluted: 1:1000). The secondary antibodies and their sources were: HRP-anti-rat IgG (Invitrogen, 31470), HRP-anti-mouse IgG (Dako, P0260), HRP-anti-rabbit IgG (Dako, P0448).

Cell sorting

For flow cytometry, cells were washed by centrifugation in PBS and resuspended in 5% (v/v) bovine calf serum (Sigma-Aldrich, F7524) in PBS. Then the cells were passed through a 50 µm filter (ctsv, 150-47S) into 5 mL tubes. Perturbations were performed as follows: For temperature, the cells were incubated at 29 °C or 39.5 °C for 16 hours prior to flow cytometry profiling. Cells were treated with 10 µM bortezomib (LC Laboratories), 20 µM chloroquine (Sigma-Aldrich), or 0.5 µM or 1 µM MLN7243 (MedChemExpress) for 16 hours prior to flow cytometry. YM01 (StressMarq Bioscience) was used at 2.5 µM and N-acetyl-aspartate (Sigma-Aldrich) at 6 mM, and added 24 hours prior to flow cytometry.

Cells were analyzed on a BD FACSJazz (BD Biosciences). Data was collected and analyzed using FlowJo (v10.7.2, BD), using the following gates: Live cells, singlet cells, BFP negative and mCherry positive.

VAMP-seq

For VAMP-seq [45], cells were grown and transfected as described above. After 5 days of treatment with doxycycline, the cells were washed by centrifugation in PBS and resuspended in 5% (v/v) bovine calf serum in PBS. Sorting was performed with a Cell Sorter BD FACS Aria III (BD Biosciences), directly based on the GFP:mCherry ratio. In total 1.1 million cells were sorted into each of four bins.

The cells were collected in tubes, pre-coated in 5% (v/v) bovine calf serum in PBS overnight, and containing 1 mL media without doxycyclin. Both the sample tube and the collecting tubes were kept at room temperature. After each sorting, the cells were harvested by centrifugation and resuspended in fresh media. The sorted cells were grown in 6-well plates for 2 days (until confluent), before being resuspended and moved to 10 cm plates to grow for another 2 days (until confluent). Next, the cells were dislodged using trypsin (0.25% (w/v) trypsin (Gibco), 10 mM Na-citrate (Sigma-Aldrich), 102.7 mM NaCl, 0.001% phenol red (Merck, 143-74-8), pH 7.8), resuspended in media, before 5 million cells from each bin were isolated and centrifuged. The supernatant was aspirated and the cell pellet stored -80 °C for later genomic DNA extraction.

Toxicity screen

Approximately 48 hours after transfection, doxycycline and AP1903 was added to the cultures as previously described. Samples of 5 million cells were taken at days 0, 5, 7 and 9 after introduction of doxycycline, as described for the VAMP-seq cells. For the day 0 sample, cells were treated with doxycyclin and AP1903 for 24 hours to select for recombinant cells, after which new media without doxycyclin was added. The cells were grown until confluent and then frozen down.

Genomic DNA extraction and sequencing

Genomic DNA extraction was performed using Qiagen DNeasy blood & tissue kit (Cat. No. 69506). Two separate purifications were performed for each sample, to be used as technical replicates in the post-sequence analysis.

For each genomic DNA sample, an adapter PCR reaction was performed as follows: All 8 tubes of a 50 µL PCR strip tube (VWR, catalog # 490003-606) were filled with 2500 ng DNA template, 25 µL 2X Q5 high fidelity Mastermix (New England Biolabs, M0492S), 0.5 µM forward primer (LC1020), 0.5 µM reverse primer (LC1031) and PCR grade H_2O to reach a total volume of 50 µL. All primers are listed in the supplemental material, (SupplementalFile2.xlsx). Samples were denatured at 98 °C (30 sec) followed by 7 cycles of PCR performed at temperatures: 98 °C (10 sec), 60 °C (20 sec), 72 °C (10 sec) and lastly a final elongation step at 72 °C (2 min). The content of the PCR tubes was pooled and mixed with an equal volume of AMPure XP beads (Beckman Coulter, A63881). After 5 min incubation, the beads were pelleted and supernatant aspirated, followed by a wash in 70 % (v/v) ethanol, and elution in 21 µL PCR grade water.

Indexing PCR reactions were performed by mixing: 4.1 µL PCR H_2O, 5 µL 5 µM forward primer, 5 µL 5 µM reverse primer, 25 µL 2X Q5 HF Mastermix, 2.5 µL 10X SYBR Green, 8.4 µL DNA template. Then, after initial denaturing at 98 °C (30 sec), 14 PCR cycles were run at the following temperatures: 98 °C (10 sec), 63 °C (20 sec), 72 °C (15 sec) followed by final elongation at 72 °C (2 min). Next, samples were mixed with DNA Gel Loading Dye (Thermo Fisher Scientific, R0611), and loaded onto a 2% agarose gel (2% (w/v) agarose in TAE-buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA with 0.01% (v/v) SYBR safe DNA gel stain (Thermo Fisher Scientific, S33102)). Electrophoresis was performed at 100 V for 50 minutes. Bands were visualized and excised using a Chemidoc imaging system (BioRad). Specific PCR product was extracted from the gel using the GeneJET gel extraction kit (Thermo Scientific, K0692) with a final elution in 30 µL PCR grade water. The DNA concentration of the eluates were measured using Qubi dsDNA High

Sensitivity (Thermo Scientific, Q32851) in order to normalize the pooled samples. The library was sequenced using a TG NextSeq 500/550 High Output Kit v2.5 (75 Cycles) (Illumina, 20024911). Custom primers were spiked in with the Illumina primers. Demultiplexing was performed using the basespace software (Illumina).

VAMP-seq data analyses

Illumina reads from the abundance and toxicity screens were cleaned for adapters using cutadapt [83] and paired-end reads were joined using fastq-join from ea-utils [84], see illumina/call_zerotol_paired.sh available on GitHub. Only barcodes with an exact match to the barcode map were counted, see illumina/merge_counts.r.

Read counts of barcodes were merged for amino acid variants and the technical replicates of each FACS bin (abundance) or time point (toxicity) and normalized to frequencies without pseudo counts. After merging, a score was calculated for variants with 20 or more reads observed per replica.

For the abundance scores, a protein stability index (PSI) was calculated per variant for each biological and FACS replica:

$$\mathrm{PSI}_i = \frac{\sum_g g \times f_{i,g}}{\sum_g f_{i,g}}$$

where $f_{i,g}$ is the frequency of variant i in FACS gate g. The reported scores are the mean and the standard deviation of PSI over replicates normalized using:

$$\text{abundance score} = \frac{\mathrm{PSI}_i - \mathrm{PSI}_{\mathrm{stop}}}{\mathrm{PSI}_{\mathrm{WT}} - \mathrm{PSI}_{\mathrm{stop}}}$$

where $\mathrm{PSI}_{\mathrm{WT}}$ is the PSI value of the wild-type amino acid sequence and $\mathrm{PSI}_{\mathrm{stop}}$ is the median PSI value of stop substitutions per amino acid residue, both averaged over all replicates, see illumine/abundance.r available on GitHub. The average Pearson correlation of scores between replica is 0.99 (range: 0.98-0.99) and mean absolute error 0.05 in normalized units (Fig. S2). All

VAMP-seq data are available on GitHub and included in the supplemental material (SupplementalFile3.xlsx).

Toxicity data analysis

The toxicity scores were calculated as the slope (α) from a linear regression of the variant frequency at each time point. Comparable slopes are achieved by normalizing the frequencies for variant, *i*, at each time point, *t*, like a distribution of frequencies:

$$\frac{f_{i,t}}{\sum_t f_{i,t}}$$

Since all sequenced pools are based on the same number of cells, we do not normalize further. This means that non-toxic variants are expected to be slightly enriched, as the complexity of the library decreases with time. We further abstain from using weighted least squares because of the inherent correlation between $f_{i,t}$ and the Poisson uncertainty of this quantity which means that the expected low frequencies of toxic variants at later time points will always be down weighted (or even ignored since pseudo counts were not applied). The reported scores are the mean and standard deviation of slopes over replicates normalized according to:

$$\text{toxicity score} = \frac{\alpha_i - \alpha_{\text{WT}}}{\alpha_{\text{C152W}} - \alpha_{\text{WT}}}$$

such that wild-type has a toxicity score of zero and the toxic C152W variant has a toxicity score of one, see illumine/toxicity.r available on GitHub. The toxicity data are available on GitHub and included in the supplemental material (SupplementalFile3.xlsx).

Degron cloning

The protein sequences of seven proteins, including ASPA, were used to construct the protein-tile library presented here and in a separate manuscript [69]. The DNA sequences of these protein sequences were optimized with the IDT codon optimization tool and then split into 72 nucleotides (nt) long oligonucleotides overlapping by 36 nt except for the C-terminal tile which may have a longer overlap. To avoid unwanted PCR products produced due to template switching over the overlapping parts of the tiles, the tiles were split into odd tiles (Odds), even tiles (Evens) and C-terminal tiles (CT) based on the position they occupy in the tile series of each protein. Two 30 nt long adaptors were attached to the 72 nt long sequences to serve as the complementary overlaps for Gibson assembly cloning resulting in 132 nt long oligos. Along with the 132 nt oligos three 126 nt long control oligos were made as well, each consisting of a 66 nt long oligo flanked by the same complementary Gibson overlaps. The three control oligos used were based on the APPY degron (-RLLL), which is 22 aa long sequence and two variants that are known to mildly (-RAAA) or strongly (-DAAA) stabilize the APPY degron [55]. The 132 nt long oligos library, and the three control oligos, were ordered from IDT as three separate libraries in a way that excludes the presence of Odds, Evens and CT oligos of the same protein in the same library tube, thus producing three libraries referred to as Odds (complexity = 93), Evens (complexity = 91) and CT (complexity = 10). The oligos were made into double-stranded DNA and amplified by the primers VV3 and VV4 using the following program: 98 °C for 30 sec and then 98 °C for 10 sec, 69 °C for 30 sec 72 °C for 10 sec for 2 cycles in total, followed by a final 72 °C incubation for 2 min. The PCR product was run on a 2% agarose gel with 1x SYBR Safe (Thermo Fisher Scientific) and the PCR product band was extracted with the GeneJet gel extraction kit (Thermo Scientific).

The Barcoded_AttB_EGFP-Link-PTEN-IRES-mCherry_Bgl2_160407 [45] vector backbone was linearized by inverse PCR with primers VV1 and VV2. The reaction was run with 5 ng of the vector DNA as template with the following program: 98 °C for 30 sec and then 98 °C for 5 sec, 69 °C for 30 sec 72 °C for 3 min and 40 sec for 30 cycles in total, followed by a final 72 °C incubation for 5 min. The PCR product was cleaned and concentrated with the Zymo Research kit following the manufacturer’s protocol and then digested by DpnI (New England BioLabs) overnight. The digestion reaction product was run on a 1% agarose gel with 1x SYBR Safe (Thermo Fisher Scientific) and the digested band was extracted from the gel with the GeneJet gel extraction kit (Thermo Scientific).

The double stranded oligos from all three libraries (Odds, Evens and CT) were assembled into the Barcoded_AttB_EGFP-Link-PTEN-IRES-mCherry_Bgl2_160407 linearized vector by a Gibson reaction using the Gibson assembly master mix (New England Biolabs) by mixing the oligos with the vector in a 4:1 molar ratio. The Gibson reaction was then cleaned and concentrated with the Zymo Research kit and transformed by electroporation into NEB 10-beta electro competent *E. coli* cells with 2kV. The electroporated cells were incubated for 1 hour at 37 °C in 1 mL and then 100 μL of a 100 fold dilution was plated on LB-ampicillin plates. The rest (900 μL) of the transformed cells were inoculated in 100 mL LB-ampicillin liquid cultures and incubated overnight. After making sure that the CFUs on the plates were at least 100x of the complexity of each library, plasmid DNA was extracted from 100 mL cultures using a midi-prep kit (Millipore Sigma) and the DNA concentration was determined by NanoDrop spectrometer ND-1000.

Tile scoring

The tiles were integrated in the HEK 293T TetBxb1BFPiCasp9 Clone 12 cell line as full-length ASPA, and sorted into 4 bins based on their GFP:mCherry ratio. DNA was extracted from the bins and amplicons were prepared for downstream Illumina high-throughput sequencing. Amplicons were amplified with primers VV40S and VV2S. The program of the first PCR (adapter PCR reaction) was the following: initial denaturation was performed at 98 °C for 30 sec; followed by 7 cycles of denaturation at 98 °C for 10 sec, annealing at 65.5 °C for 10 sec and extension at 72 °C for 50 sec; a final extension at 72 °C for 2 min. Afterwards, the product was purified by Ampure XP beads (Beckman Coulter) (0.8:1 ratio) (beads: PCR reaction product) and the Illumina cluster generation sequences were added with a second PCR (indexing PCR reaction) with the primers gDNA_2nd and JS_R. The PCR program used for the second PCR is as follows: initial denaturation at 98 °C for 30 sec; followed by 16 cycles of denaturation at 98 °C for 10 sec, annealing at 63.5 °C for 10 sec and extension at 72 °C for 10 sec. The amplicons were sequenced by a NextSeq 550 sequencer with a NextSeq 500/550 Mid Output v2.5 300 cycle kit (Illumina) with custom sequencing primers VV16 and VV18 for read 1 and read 2 (paired-end). The indices were read with the primers VV19 and VV21 for index 1 and index 2 respectively.

Similar to the processing of reads in the VAMP-seq experiment, the tile reads were cleaned for adapters sequences using cutadapt [83] and paired end reads were joined using fastq-join from ea-utils [84]. Only barcodes with an exact match to the barcode map were counted. If tiles from the Odds, Evens or CT libraries were observed in a sorting of a different library, these were assumed to be non-sorted contaminants and ignored. Technical replicates of each FACS bin were merged and normalized to frequencies without pseudo counts. For each library, biological and FACS replicas, a tile stability index (TSI) was calculated per tile using:

$$\mathrm{TSI}_t = \frac{\sum_g g \times f_{t,g}}{\sum_g f_{t,g}}$$

Where $f_{t,g}$ is the frequency of tile t in FACS gate g. Two of the APPY based control tiles, RLLL and DAAA, present in all sequenced pools were used to renormalize the Evens and CT libraries to match the TSI of the control tiles to the Odds library:

$$\mathrm{TSI}_t^{\mathrm{even,norm}} = 0.075 + 0.9018 * \mathrm{TSI}_t^{\mathrm{even}}$$

$$\mathrm{TSI}_t^{\mathrm{ct,norm}} = 0.5895 + 0.5570 * \mathrm{TSI}_t^{\mathrm{ct}}$$

Since the complexity of the libraries is relatively low, each tile is covered by more than 3500 observed reads per technical replicate on average. Thus 3 biological and 2 FACS replicates for each of the 3 libraries reproduced TSI scores very well with a minimum Pearson correlation of 0.97. The standard deviation over replicates is reported as error estimate. The tile scores are available on the

GitHub repository of this paper and the code is available with the original report of this experiment[69].

Evolutionary conservation scores

Evolutionary distance from WT sequence were calculated *in silico* for all the ASPA variants using information from evolutionary sequence conservation. We generated a multiple sequence alignment (MSA) of ASPA homologs using HHblits [85] with an E-value threshold of 10^{-20}. The full ASPA MSA included 1102 sequences, but was reduced to 757 homologs by filtering out sequences with more than 50% gaps. Using the MSA information, we calculated evolutionary conservation scores using the Global Epistatic Model for predicting Mutational Effects (GEMME) software [51].

In silico thermodynamic stability predictions

Changes in thermodynamic stability (ΔΔG) were predicted using Rosetta (GitHub SHA1 99d33ec59ce9fcecc5e4f3800c778a54afdf8504) with the Cartesian ddG protocol [50] on ASPA crystal structures 2O53, using only the chain A for the monomeric evaluation and both the chains (AB) for the dimeric evaluation. Non-protein atoms were removed from the crystal structure except for the zinc ion that was kept in the dimer calculations. All the ΔΔG values obtained from Rosetta were divided by 2.9 to bring them from Rosetta energy units onto a scale corresponding to kcal/mol [50] and truncated to the range 0-5 kcal/mol.

RNA sequencing and qPCR

After 5 days of induced protein expression using doxycycline, viable, singlet, mCherry-positive cells were sorted using a BD FACSJazz (BD Biosciences).

For the qPCR, RNA was purified from samples containing more than 500,000 cells using an RNeasy kit (Qiagen) following the protocol without the optional on-column DNase digestion step. DNA digestion was performed on 1 µg nucleic acid using DNase I, RNase-free (ThermoFisher Scientific) as described by the manufacturer. 1 U/µL RiboLock RNase Inhibitor (Thermo Scientific), was included to prevent RNA degradation. Subsequently, reverse transcription was performed using Maxima H Minus Reverse Transcriptase (Thermo Scientific), yielding 20 µL cDNA sample. Of these, 1 µL was mixed with 12.5 µL Maxima SYBR Green/ROX qPCR Master Mix (2X) (Thermo Scientific), primers at final concentration of 0.3 µM for either HSPA1A/B or β-actin and water for a final volume of 25 µL. The samples were denatured at 95 °C (10 min), followed by 40 cycles performed at the temperatures: 95 °C (15 sec), 59 °C (30 sec), 72 °C (30 sec). The mean of 3 replicates (not differing by more than 0.5 cycle) was calculated and normalized to β-actin. Subsequently samples were normalized to a WT sample included in each PCR run. In total, 3 biological replicates were included per ASPA variant.

For the RNA sequencing, total RNA was isolated and purified from more than 1.5 million sorted cells using the GeneJET RNA Purification Kit (ThermoFisher Scientific) according to the manufacturer's instruction. Genomic DNA was removed from 5 µg total RNA according to the manufacturer's instruction in the GeneJET RNA Purification Kit (ThermoFisher Scientific). RNA samples of more than 2.5 µg were shipped to BGI (Hong Kong). The RNA was rRNA depleted and sequenced using DNBSEQ by BGI.

RNA sequencing data analysis

Quality control of sequence reads was done using the tools "FastQC" v0.11.7 (), "RSeQC" v2.6.4 [86] and "fastq_screen" v0.11.4 (https: //www.bioinformatics.babraham.ac.uk/projects/fastq_screen/). Low-quality bases and the first 12 bases and reads shorter than 25 nt were removed with

“Trimmomatic” v0.39 [87] using settings “HEADCROP:12 LEADING:3 SLIDINGWINDOW:4:15 MINLEN:35”. Reads were mapped using “STAR” v2.7.3a [88] against the human genome (hg38). Up to two mismatches were allowed during the mapping, and the minimum number of overlap bases to trigger mates merging and realignment was set to five. Otherwise, default settings were used. Duplicate reads were also removed with the bamRemoveDuplicatesType "UniqueIdentical" option in “STAR”. The “featureCounts” function of the “Rsubread” R package v2.2.6 [89] was used to quantify reads in exons. The Gencode v38 comprehensive gene annotation including all genomic regions was used to assign reads to genes.

The “edgeR” v3.30.6 software [90] was used to perform a differential expression analysis. For this purpose, first, a model was defined indicating the experimental conditions. Library normalization factors were calculated using the “calcNormFactors” function with the “TMM” algorithm. Tag-wise dispersion was calculated using the “estimateDisp” function with “robust = TRUE”. A gene-wise generalized linear model was fit with “glmQLFit”. Finally, differential gene usage was assessed using “glmQLFTest”. Resulting p-values were corrected for multiple testing using the “Benjamini-Hochberg” method.

Acknowledgements

We acknowledge the use of the FACS, microscopy, bioinformatics and computing core facilities at the Biotech Research & Innovation Centre and the Department of Biology, University of Copenhagen. We thank Vibe H. Oestergaard, Michael Lisby, Leonor Rib, Nick Popp, Sofie V. Nielsen, Søren Lindemose and Anne-Marie Lauridsen for technical support.

Competing interests

K.L.-L. holds stock options in and is a consultant for Peptone Ltd.

Data and availability

All data and software generated for this article is available on GitHub: https://github.com/KULL-Centre/_2023_Groenbaek-Thygesen_ASPA_MAVE (DOI:10.5281/zenodo.8382504). Abundance and toxicity scores are also deposited at MaveDB (https: //www.mavedb.org) under accession number urn:mavedb:00000657-a. Sequencing reads for the abundance and toxicity scores are available at https: //doi.org/10.17894/ucph.3e05fe3a-4d7e-4d70-9056-18ed999e7e1e. The RNA seq. data have been uploaded to Gene Expression Omnibus (GEO): https: //www.ncbi.nlm.nih.gov/geo/ (accession number: GSE232399; samples GSM7329952-57).

Supplemental files

This article includes the following supplemental information.

- Supplemental Figures and Tables
- Abundance, toxicity, Rosetta & GEMME for selected variants (SupplementalFile1.xlsx)
- Sequences of primers (SupplementalFile2.xlsx)
- All abundance and toxicity screening data (SupplementalFile3.xlsx)
- Results from the RNA sequencing (SupplementalFile4.xlsx)

Author contributions

M.G.-T., V.V., K.E.J., M.C., L.P., T.K.S., L.C., S.N., R.L.P. and A.S. performed the experiments. M.G.-T., V.V., M.C., K.E.J., T.K.S., L.C., A.S., D.M.F., K.L.-L. and R.H.-P. analyzed the data. D.M.F., K.L.-L. and R.H.-P. conceived the study. M.G.-T., V.V and R.H.-P. wrote the paper.

Funding

The work was funded by the Novo Nordisk Foundation (https:// novonordiskfonden.dk) challenge program PRISM (to K.L.-L., A.S., D.M.F. & R.H.-P.), the Lundbeck Foundation (https://www.lundbeckfonden.com) R272-2017-452 and R209-2015-3283 (to A.S.), and R249-2017-510 (to L.C.). Danish Council for Independent Research (Det Frie Forskningsråd) (https://dff.dk) 10.46540/2032-00007B (to R.H.P.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Kwon, Y. T. & Ciechanover, A. The Ubiquitin Code in the Ubiquitin-Proteasome System and Autophagy. *Trends Biochem. Sci.* **42**, 873–886 (2017).
2. Klaips, C. L., Jayaraj, G. G. & Hartl, F. U. Pathways of cellular proteostasis in aging and disease. *J. Cell Biol.* **217**, 51–63 (2018).
3. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).
4. Chhangani, D., Joshi, A. P. & Mishra, A. E3 ubiquitin ligases in protein quality control mechanism. *Mol. Neurobiol.* **45**, 571–585 (2012).
5. Lapidus, L. J. Protein unfolding mechanisms and their effects on folding experiments. *F1000Research* **6**, 1723 (2017).
6. Anfinsen, C. B., HABER, E., SELA, M. & WHITE, F. H. J. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 1309–1314 (1961).
7. Taipale, M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms (1). *Biochem. Cell Biol.* **97**, 46–57 (2019).
8. Casadio, R., Vassura, M., Tiwari, S., Fariselli, P. & Luigi Martelli, P. Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* **32**, 1161–1170 (2011).
9. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
10. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *Elife* **12**, (2023).
11. Hernández-Ramírez, L. C. *et al.* Rapid Proteasomal Degradation of Mutant Proteins Is the Primary Mechanism Leading to Tumorigenesis in Patients With Missense AIP Mutations. *J. Clin. Endocrinol. Metab.* **101**, 3144–3154 (2016).
12. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends Biochem. Sci.* **44**, 575–588 (2019).
13. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022).
14. Cagiada, M. *et al.* Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol. Biol. Evol.* **38**, 3235–3246 (2021).
15. Sanavia, T. *et al.* Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput. Struct. Biotechnol. J.* **18**, 1968–1979 (2020).
16. Pancotti, C. *et al.* Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* **23**, 1–12 (2022).
17. Clark, W. T. *et al.* Assessment of predicted enzymatic activity of α-N-acetylglucosaminidase variants of unknown significance for CAGI 2016. *Hum. Mutat.* **40**, 1519–1529 (2019).
18. Kirmani, B. F., Jacobowitz, D. M. & Namboodiri, M. A. A. Developmental increase of aspartoacylase in oligodendrocytes parallels CNS myelination. *Dev. Brain Res.* **140**, 105–115 (2003).
19. Klugmann, M. *et al.* Identification and distribution of aspartoacylase in the postnatal rat brain. *Neuroreport* **14**, 1837–1840 (2003).
20. Madhavarao, C. N. *et al.* Immunohistochemical Localization of Aspartoacylase in the Rat Central Nervous System. *J. Comp. Neurol.* **472**, 318–329 (2004).
21. Schuff, N. *et al.* N-acetylaspartate as a marker of neuronal injury in neurodegenerative disease. *Adv. Exp. Med. Biol.* **576**, 241–262 (2006).
22. Kaul, R., Gao, G. P., Balamurugan, K. & Matalon, R. Cloning of the human aspartoacylase cDNA and a common missense mutation in Canavan disease. *Nat. Genet.* **5**, 118–123 (1993).
23. Bitto, E., Bingman, C. A., Wesenberg, G. E., McCoy, J. G. & Phillips, G. N. Structure of

aspartoacylase, the brain enzyme impaired in Canavan disease. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 456–461 (2007).

24. Le Coq, J. *et al.* Examination of the Mechanism of Human Brain Aspartoacylase through the Binding of an Intermediate Analogue. *Biochemistry* **47**, 3484–3492 (2008).

25. Hershfield, J. R. *et al.* Aspartoacylase is a regulated nuclear□cytoplasmic enzyme. *FASEB J.* **20**, 2139–2141 (2006).

26. Moore, R. A., Le Coq, J., Faehnle, C. R. & Viola, R. E. Purification and preliminary characterization of brain aspartoacylase. *Arch. Biochem. Biophys.* **413**, 1–8 (2003).

27. Le Coq, J., An, H.-J., Lebrilla, C. & Viola, R. E. Characterization of Human Aspartoacylase: the brain enzyme responsible for Canavan disease. *Biochemistry* **46**, 5878–5884 (2006).

28. Baslow, M. H. Canavan's spongiform leukodystrophy: A clinical anatomy of a genetic metabolic CNS disease. *J. Mol. Neurosci.* **15**, 61–69 (2000).

29. Baslow, M. H. & Guilfoyle, D. N. Canavan disease, a rare early-onset human spongiform leukodystrophy: Insights into its genesis and possible clinical interventions. *Biochimie* **95**, 946–956 (2013).

30. D'Adamo, A. F. & Yatsu, F. M. ACETATE METABOLISM IN THE NERVOUS SYSTEM. N□ACETYL□l□ASPARTIC ACID AND THE BIOSYNTHESIS OF BRAIN LIPIDS. *J. Neurochem.* **13**, 961–965 (1966).

31. Matalon, R. & Michals-Matalon, K. Molecular basis of Canavan disease. *Eur. J. Paediatr. Neurol.* **2**, 69–76 (1998).

32. Matalon, R. & Michals-Matalon, K. Biochemistry and molecular biology of Canavan disease. *Neurochem. Res.* **24**, 507–513 (1999).

33. Hoshino, H. & Kubota, M. Canavan disease: Clinical features and recent advances in research. *Pediatr. Int.* **56**, 477–483 (2014).

34. Roscoe, R. B., Elliott, C., Zarros, A. & Baillie, G. S. Non-genetic therapeutic approaches to Canavan disease. *J. Neurol. Sci.* **366**, 116–124 (2016).

35. Topçu, M. *et al.* Effect of topiramate on enlargement of head in Canavan disease: a new option for treatment of megalencephaly. *Turk. J. Pediatr.* **46**, 67–71 (2004).

36. Miranda, C. O., Brites, P., Sousa, M. M. & Teixeira, C. A. Advances and pitfalls of cell therapy in metabolic leukodystrophies. *Cell Transplant.* **22**, 189–204 (2013).

37. Nešuta, O. *et al.* High Throughput Screening Cascade to Identify Human Aspartate N-Acetyltransferase (ANAT) Inhibitors for Canavan Disease. *ACS Chem. Neurosci.* (2021) doi:10.1021/acschemneuro.1c00455.

38. Edo Solsona, M. D., Fernández, L. L., Boquet, E. M. & Andrés, J. L. P. Lithium citrate as treatment of canavan disease. *Clin. Neuropharmacol.* **35**, 150–151 (2012).

39. Janson, C. *et al.* Clinical protocol. Gene therapy of Canavan disease: AAV-2 vector for neurosurgical delivery of aspartoacylase gene (ASPA) to the human brain. *Hum. Gene Ther.* **13**, 1391–1412 (2002).

40. McPhee, S. W. J. *et al.* Immune responses to AAV in a phase I study for Canavan disease. *J. Gene Med.* **8**, 577–588 (2006).

41. Leone, P. *et al.* Long-term follow-up after gene therapy for canavan disease. *Sci. Transl. Med.* **4**, 165ra163-165ra163 (2012).

42. Gray, S. J. Timing of Gene Therapy Interventions: The Earlier, the Better. *Mol. Ther.* **24**, 1017–1018 (2016).

43. Ahmed, S. S. *et al.* A single intravenous rAAV injection as late as P20 achieves efficacious and sustained CNS Gene therapy in Canavan mice. *Mol. Ther.* **21**, 2136–2147 (2013).

44. Gersing, S. K. *et al.* Mapping the degradation pathway of a disease-linked aspartoacylase variant. *PLoS Genet.* **17**, 1–28 (2021).

45. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).

46. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* **48**, 1–12 (2020).

47. Cagiada, M. *et al.* Discovering functionally important sites in proteins. *Nat. Commun.* **14**, 4175

(2023).

48. Abildgaard, A. B. *et al.* Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. *Elife* **8**, e49138 (2019).
49. Nielsen, S. V. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet.* **13**, 1–26 (2017).
50. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
51. Laine, E., Karami, Y. & Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol. Biol. Evol.* **36**, 2604–2619 (2019).
52. Geffen, Y. *et al.* Mapping the Landscape of a Eukaryotic Degronome. *Mol. Cell* **63**, 1055–1065 (2016).
53. Maurer, M. J. *et al.* Degradation Signals for Ubiquitin-Proteasome Dependent Cytosolic Protein Quality Control (CytoQC) in Yeast. *G3 (Bethesda).* **6**, 1853–1866 (2016).
54. Kampmeyer, C. *et al.* Disease-linked mutations cause exposure of a protein quality control degron. *Structure* **30**, 1245-1253.e5 (2022).
55. Abildgaard, A. B. *et al.* HSP70-binding motifs function as protein quality control degrons. *Cell. Mol. Life Sci.* **80**, 32 (2023).
56. Johansson, K. E., Mashahreh, B., Hartmann-Petersen, R., Ravid, T. & Lindorff-Larsen, K. Prediction of quality-control degradation signals in yeast proteins. *J. Mol. Biol.* **Volume 435**, (2023).
57. Mashahreh, B. *et al.* Conserved degronome features governing quality control-associated proteolysis. *Nat. Commun.* **13**, (2022).
58. Timms, R. T. & Koren, I. Tying up loose ends: the N-degron and C-degron pathways of protein degradation. *Biochem. Soc. Trans.* **48**, 1557–1567 (2020).
59. Koren, I. *et al.* The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. *Cell* **173**, 1622-1635.e14 (2018).
60. Guharoy, M., Bhowmick, P., Sallam, M. & Tompa, P. Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. *Nat. Commun.* **7**, (2016).
61. Inobe, T., Fishbain, S., Prakash, S. & Matouschek, A. Defining the geometry of the two-component proteasome degron. *Nat. Chem. Biol.* **7**, 161–167 (2011).
62. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
63. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
64. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science* **353**, aac4354 (2016).
65. Kepp, K. P. A quantitative model of human neurodegenerative diseases involving protein aggregation. *Neurobiol. Aging* **80**, 46–55 (2019).
66. Kubota, H. Quality Control Against Misfolded Proteins in the Cytosol: A Network for Cell Survival. *J. Biochem.* **146**, 609–616 (2009).
67. Sakahira, H., Breuer, P., Hayer-Hartl, M. K. & Hartl, F. U. Molecular chaperones as modulators of polyglutamine protein aggregation and toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 4**, 16412–16418 (2002).
68. Chiti, F. & Dobson, C. M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).
69. Clausen, L. *et al.* A mutational atlas for Parkin proteostasis. *bioRxiv* 2023.06.08.544160 (2023) doi:10.1101/2023.06.08.544160.
70. VanPelt, J. & Page, R. C. Unraveling the CHIP:Hsp70 complex as an information processor for protein quality control. *Biochim. Biophys. acta. Proteins proteomics* **1865**, 133–141 (2017).
71. Edkins, A. L. CHIP: a co-chaperone for degradation by the proteasome. *Subcell. Biochem.* **78**, 219–242 (2015).
72. McDonough, H. & Patterson, C. CHIP: a link between the chaperone and proteasome systems. *Cell Stress Chaperones* **8**, 303–308 (2003).

73. Theodoraki, M. A., Nillegoda, N. B., Saini, J. & Caplan, A. J. A network of ubiquitin ligases is important for the dynamics of misfolded protein aggregates in yeast. *J. Biol. Chem.* **287**, 23911–23922 (2012).
74. Samant, R. S., Livingston, C. M., Sontag, E. M. & Frydman, J. Distinct proteostasis circuits cooperate in nuclear and cytoplasmic protein quality control. *Nature* **563**, 407–411 (2018).
75. Breckel, C. A. & Hochstrasser, M. Ubiquitin Ligase Redundancy and Nuclear-Cytoplasmic Localization in Yeast Protein Quality Control. *Biomolecules* **11**, (2021).
76. Arlow, T., Scott, K., Wagenseller, A. & Gammie, A. Proteasome inhibition rescues clinically significant unstable variants of the mismatch repair protein Msh2. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 246–251 (2013).
77. Wilke, M., Bot, A., Jorna, H., Scholte, B. J. & de Jonge, H. R. Rescue of murine F508del CFTR activity in native intestine by low temperature and proteasome inhibitors. *PLoS One* **7**, e52070 (2012).
78. Roda, J. *et al.* New drugs in cystic fibrosis: what has changed in the last decade? *Ther. Adv. Chronic Dis.* **13**, 20406223221098136 (2022).
79. Wei, H. *et al.* The pathogenesis of, and pharmacological treatment for, Canavan disease. *Drug Discov. Today* **27**, 2467–2483 (2022).
80. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.* **137**, 665–678 (2018).
81. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
83. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
84. Aronesty, E. Comparison of Sequencing Utility Programs. *Open Bioinforma. J.* **7**, 1–8 (2013).
85. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
86. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
87. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
88. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
89. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
90. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

Legends to figures

Fig. 1. *The ASPA expression system.* (A) Schematic representation of the expression system. HEK293T cells, carrying a landing pad for Bxb1-catalyzed site-specific integration are transfected with the expression vector and a Bxb1 expression plasmid (not shown). Upon integration at the landing pad locus, the BFP-iCasp9-Blast^R gene is displaced downstream, and the cells therefore become resistant to AP1903, while GFP-ASPA and mCherry is expressed from the tetracycline/doxycyclin regulated promoter. The same mRNA leads to both GFP-ASPA and mCherry protein production, which in turn allows flow sorting of cells based on the GFP:mCherry ratio. Finally, variants in each bin can be identified by sequencing the barcodes. (B) Fluorescent microscopy of cells transfected with either wild-type ASPA (WT) or ASPA C152W variant. Note the reduced amount of the C152W variant. Scale bar = 20 µm. (C) Cells were transfected with either WT or C152W ASPA variants fused to GFP in the N-terminus or C-terminus as indicated. A mock transfection was included as a control. Whole-cell lysates were then resolved by SDS-PAGE and analyzed by Western blotting using antibodies to GFP, mCherry or, as a loading control, GAPDH. Note the reduced level of the C152W variant. (D) Scatterplots of flow cytometry analyses of the WT (blue) and C152W (red) ASPA variants, along with the site-saturated ASPA library (grey). Note that the mCherry levels are similar, while the GFP levels differ approximately 10-fold. (E) Histograms of the GFP:mCherry ratio based of WT (blue) and C152W (red) ASPA variants, and the ASPA variant library (grey). (F) The ASPA library was sorted into four separate bins (1-4) as indicated, with each bin containing 25% of the total population.

Fig. 2. *The ASPA abundance map*. (A) The results of the ASPA abundance screen described in Fig. 1A are presented as a heat map with the position in ASPA (horizontal) and the 20 different amino acids (vertical). * indicates a stop codon. The median abundance score (MED) per position is shown above. The wild-type residues are shown in yellow. Missing data points are marked in grey. Neutral variants (WT-like abundance) are in white. Low abundance variants are shown in red and high abundance variants are shown in blue. For comparison, the AlphaFold confidence scores (pLDDT) are marked below. Regions with low pLDDT scores (green/blue colors) indicate flexible/disordered regions. The ASPA domain organization and secondary structure are marked. (B) The library displays a bimodal distribution of abundance scores with a peak of neutral variants overlapping with the synonymous (silent) WT ASPA variants, and a peak of low abundance variants overlapping with the nonsense ASPA variants. (C) To validate the abundance map, 18 ASPA variants were generated and analyzed one-by-one by flow cytometry in low-throughout. The abundance scores determined in low-throughput (y-axis) correlate with the abundance scores determined from the screen (x-axis). (D) The ASPA dimer structure (PDB: 2O53) colored by the median abundance score. The Zn^{2+} ions are marked as yellow spheres. Note that the surface of ASPA appears more tolerant to amino acid substitutions than regions that are buried or located in the subunit-subunit interface.

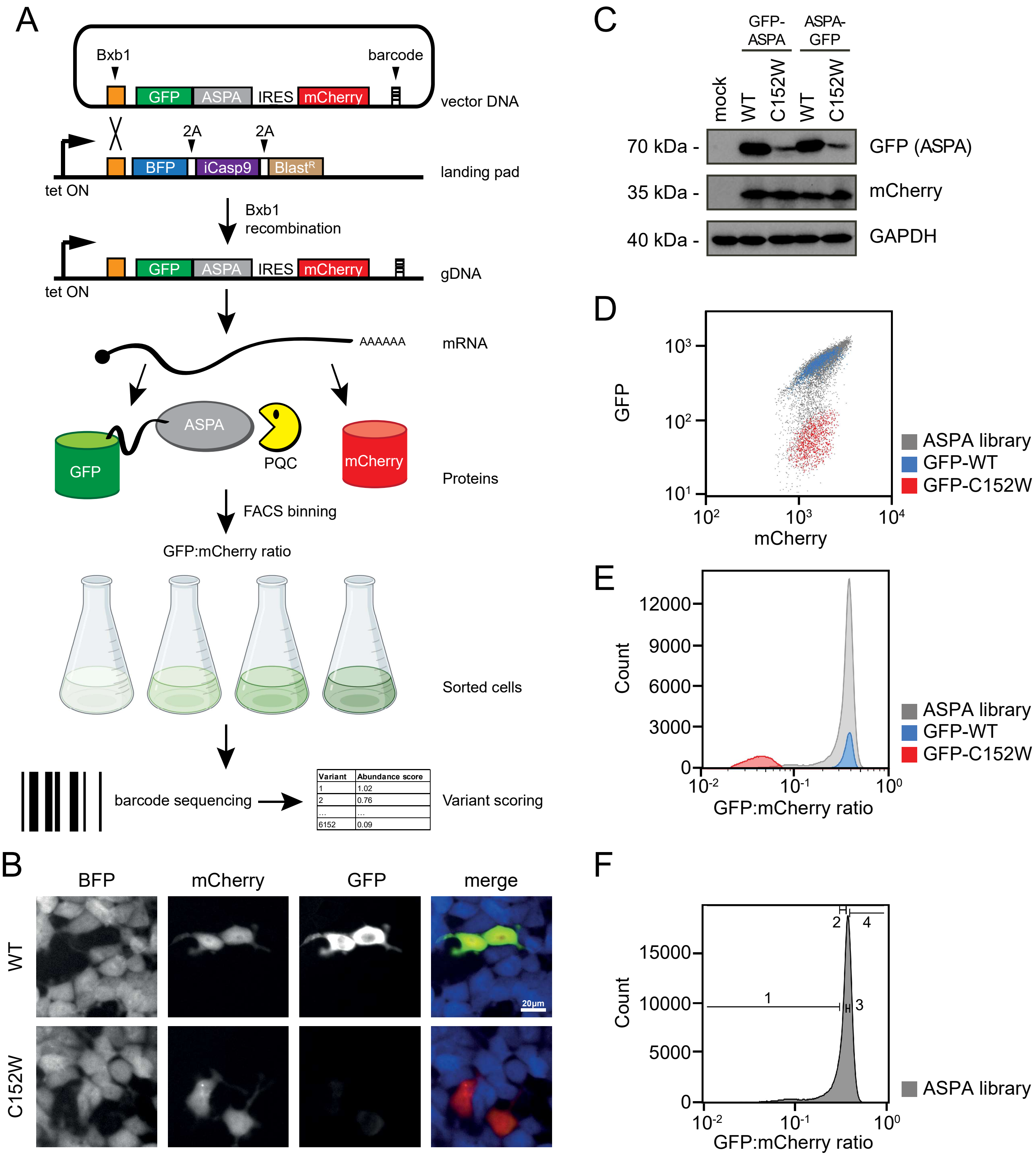
Fig. 3. *Correlations with thermodynamic stability predictions and evolutionary conservation.* (A) Scatter plots showing correlations between the abundance scores and the predicted protein stabilities (ΔΔG) for all variants (left panel) and the median scores per position (right panel). (B) Scatter plots showing correlations between the abundance scores and the evolutionary conservation scores for all variants (left panel) and the median scores per position (right panel). CI, bootstrapped 95% confidence interval.

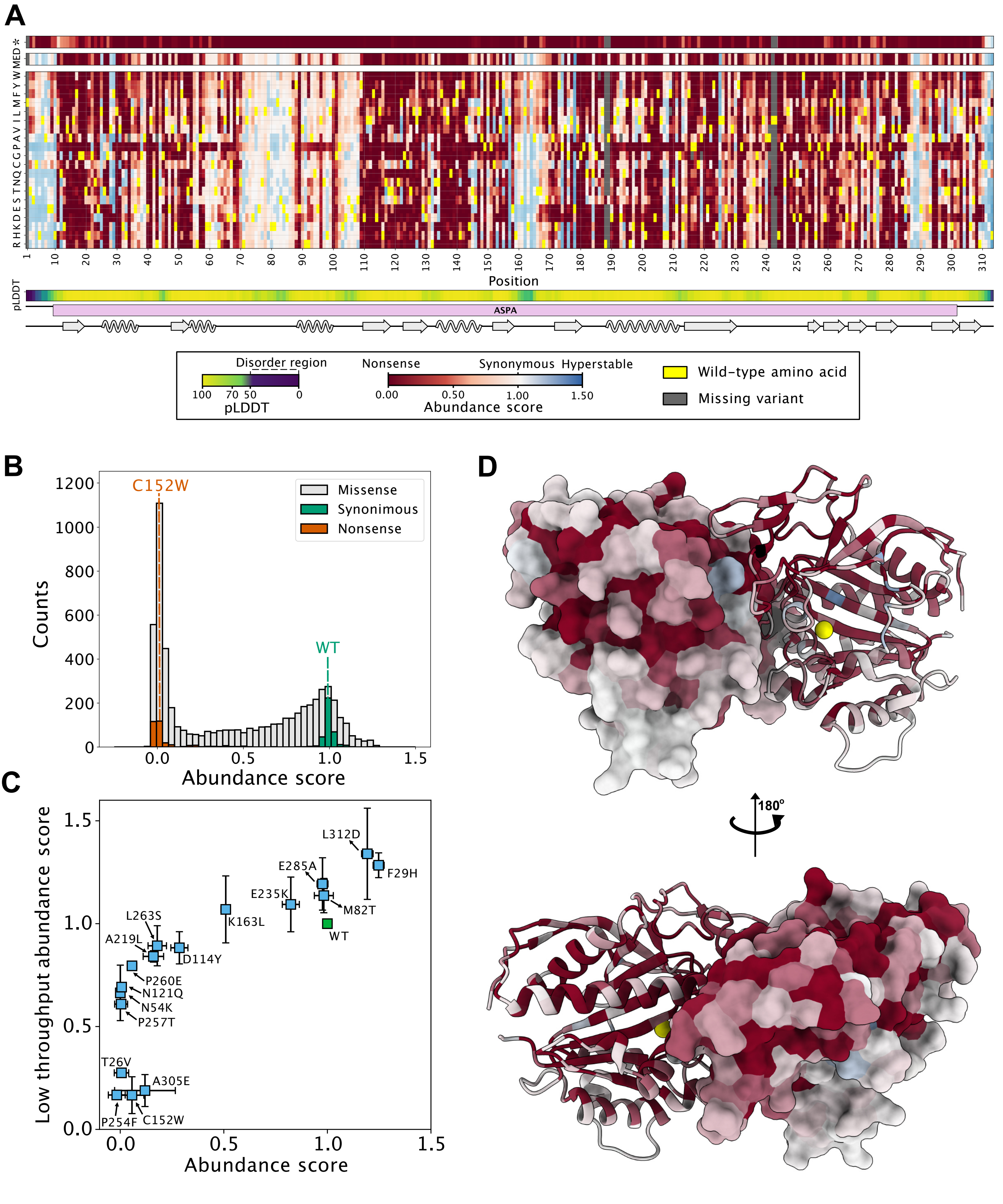
Fig. 4. *Flow cytometry distributions of the ASPA library with different cellular perturbations.* Histograms displaying the distributions of the GFP:mCherry ratios of the ASPA library, and for comparison ASPA WT and C152W (A), were analyzed for the indicated perturbations: (B) 16 hours incubation at 39.5 °C, (C) 16 hours incubation at 29 °C, (D) 16 hours with 0.5 μM (blue) or 1 μM (red) of the ubiquitin E1-inhibitor MLN7243, (E) 16 hours with 15 μM of the proteasome inhibitor bortezomib (BZ), (F) 16 hours with 20 μM the lysosomal inhibitor chloroquine (CQ), (G) 24 hours with 2.5 μM of the HSP70 –inhibitor YM01, and (H) 24 hours with 6 mM N-acetyl-aspartate (NAA).

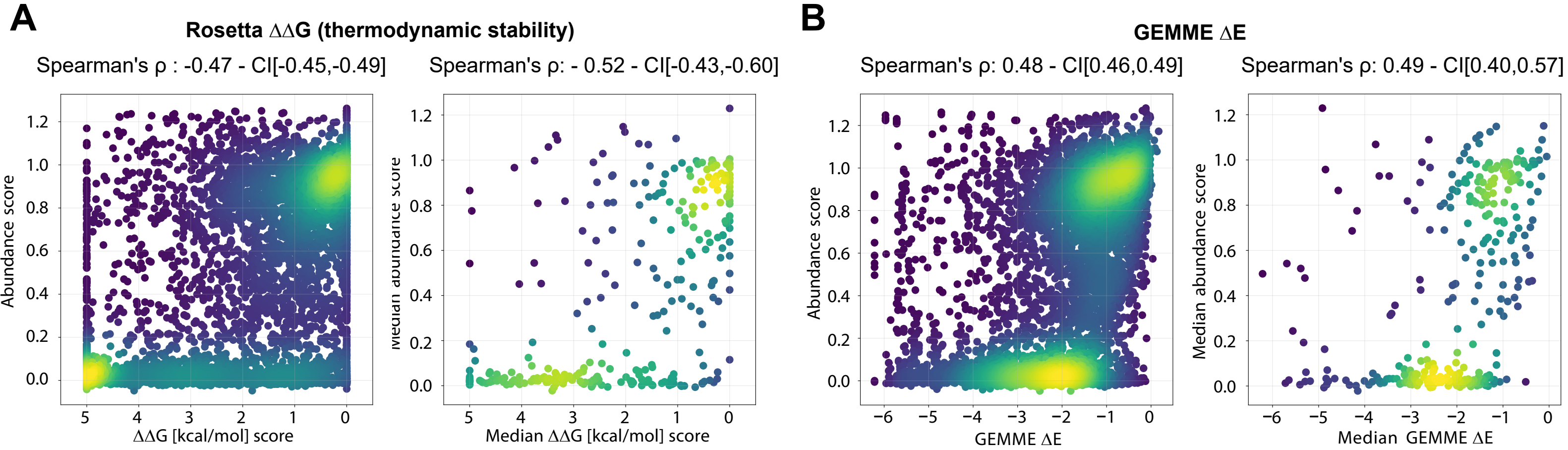
Fig. 5. *Mapping inherent degrons in ASPA.* (A) The ASPA protein was divided into 26 different tiles of 24-residues, each overlapping by 12 residues, as indicated. (B) The ASPA tiles shown in panel A were expressed from the landing pad in HEK293T cells. Then the cells were flow sorted into different bins based on GFP:mCherry ratio and the tiles in each bin were identified by sequencing across the tiles. (C) The sequencing shown in panel B was used to determine a tile stability index (TSI) for each of ASPA tiles. Each point is positioned at the central position of the 24-mer tiles. Tiles with a low TSI have reduced GFP:mCherry ratios and therefore display degron-like properties. As a measure for exposure, the weighted contact number (WCN) was determined for each tile based on the ASPA crystal structure (PDB: 2O53). The domain organization of ASPA is included for comparison. (D) PQC degrons in ASPA were predicted from the ASPA sequence using QCDPred. Note that regions where QCDPred predicts a high probability of PQC degrons overlap with regions with a low TSI (panel C).

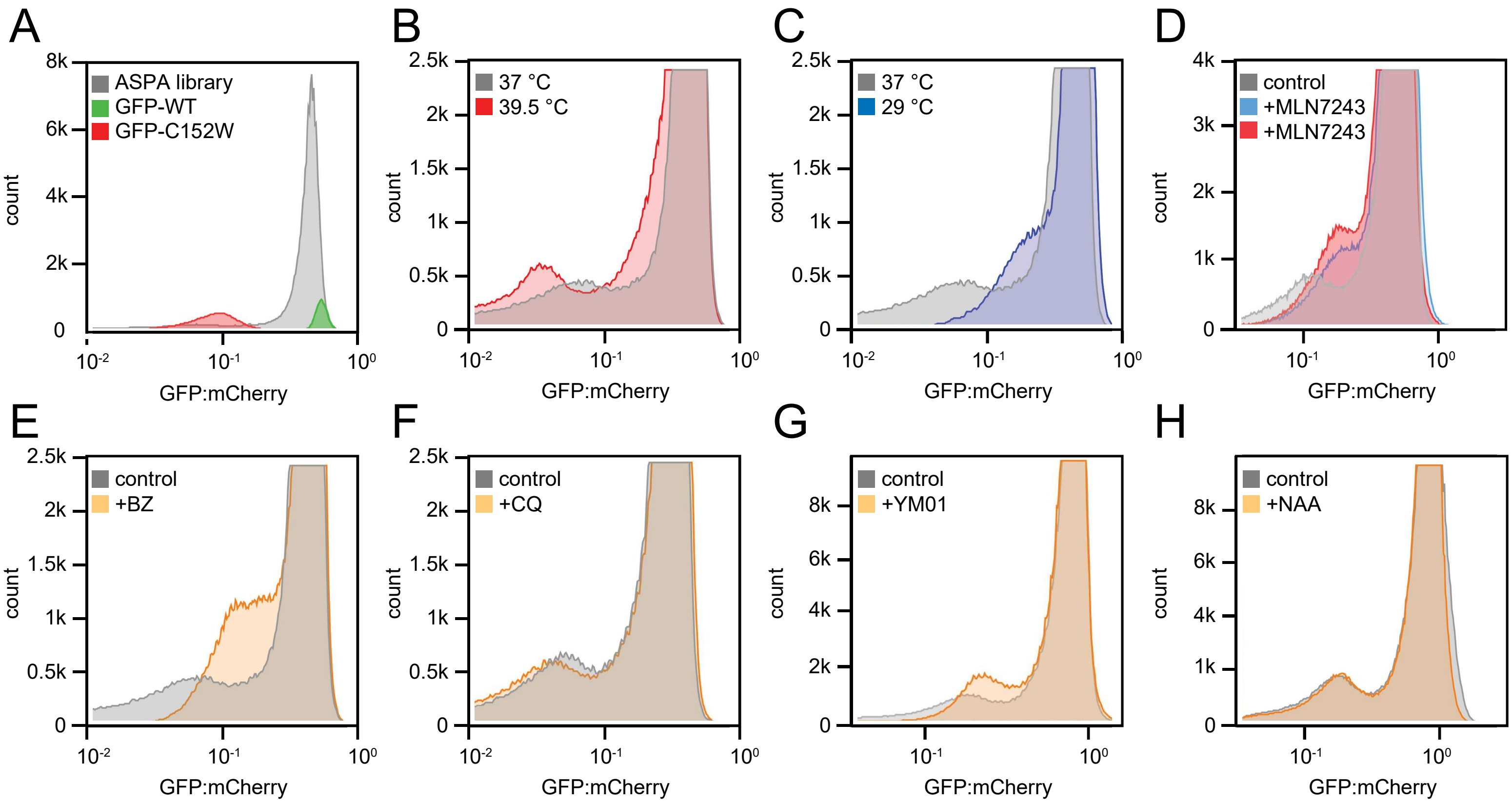
Fig. 6. *Comparing the ASPA abundance scores with human genetics data.* (A) Comparisons of the abundance scores for ASPA missense variants listed in Supplemental File 1 as pathogenic (red), variants of uncertain significance (VUS) (yellow) and benign (green) are shown as raincloud plots. Residues in or near the ASPA active site have been marked (blue). Note that many pathogenic and some VUS variants display a low abundance. Many of the high abundance pathogenic variants are located near the active site. (B) Comparison of the ASPA abundance scores with the ASPA allele frequency reported in gnomAD. Note that ASPA variants that are common in the population are benign and display a wild-type-like abundance, while many rare variants display a low abundance. Variants so rare that they have not been observed in gnomAD are included to the left of the dashed line. (C) Comparison of the abundance scores with GEMME evolutionary conservation scores. All variants are shown as a blue 2D histogram and overlayed with variants annotated as pathogenic (red), benign (green) and VUS (yellow).

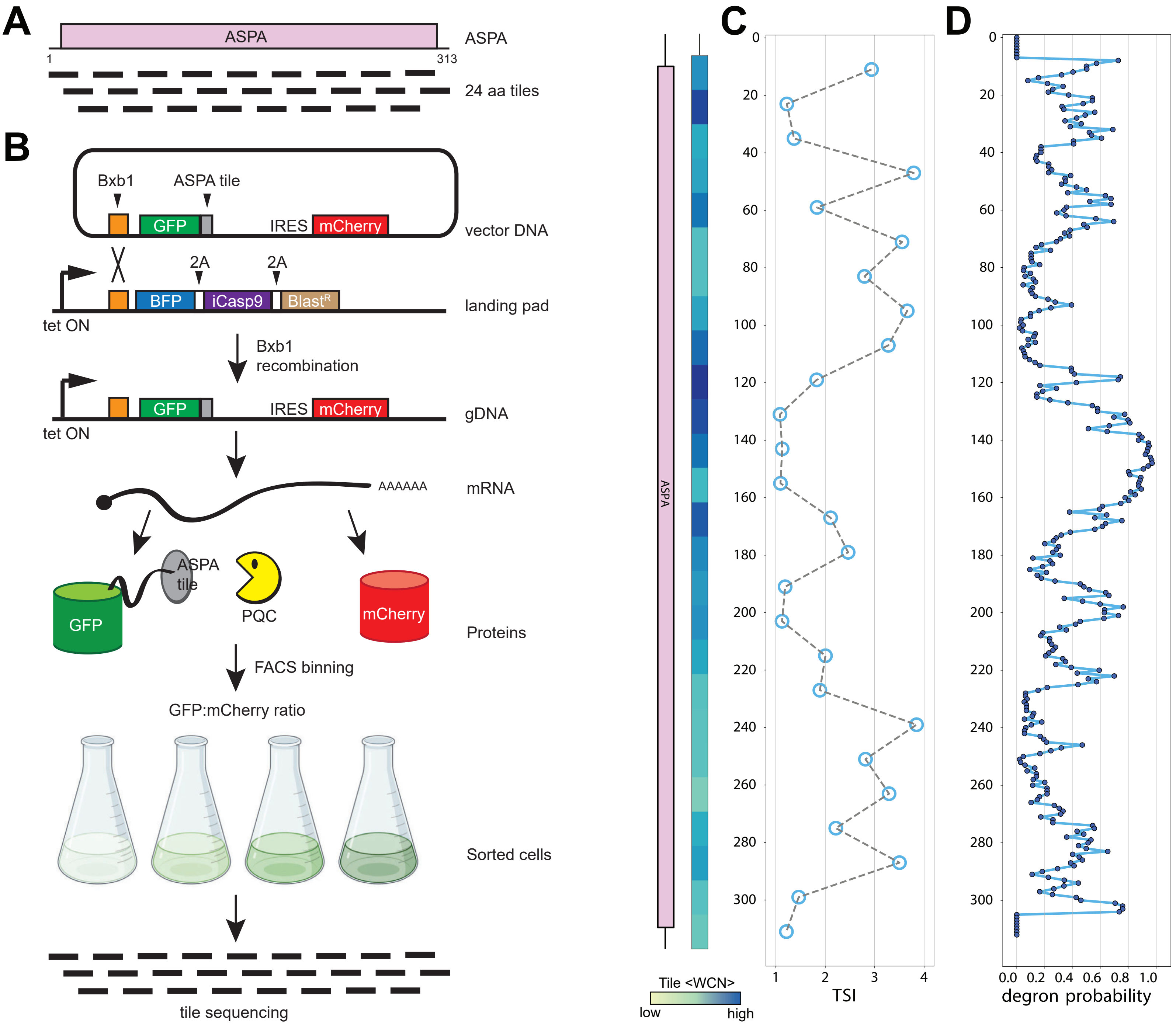
Fig. 7. *Some low abundant ASPA variants are toxic.* (A) Results from ASPA toxicity screen presented as a heat map with the position in ASPA (horizontal) and the 20 different amino acids (vertical). * indicates a stop codon. The median toxicity score (MED) per position is shown above. The wild-type residues are shown in yellow. Missing data points are marked in grey. Non-toxic variants (WT-like) are in white. Toxic variants are shown in green. (B) Correlation between abundance and toxicity scores for all missense variants. Note that all toxic variants have low abundance scores. (C) Plot showing correlation between toxicity and Rosetta ΔΔG values for all missense variants. (D) Plot showing correlation between toxicity and GEMME scores for all missense variants. In panels B, C and D, the correlations are illustrated using 2D histograms consisting of hexagonal bins, with the number of data points in each hexagon determining the color of the bin. The data point densities are shown according to the color scales below each individual plot. CI, bootstrapped 95% confidence interval. (E) The ASPA dimer structure (PDB: 2O53) colored by the median toxicity score. The Zn^{2+} ions are marked as yellow spheres. Note that toxicity is most pronounced in amino acid substitutions within regions that are buried or located in the subunit-subunit interface compared to the surface.

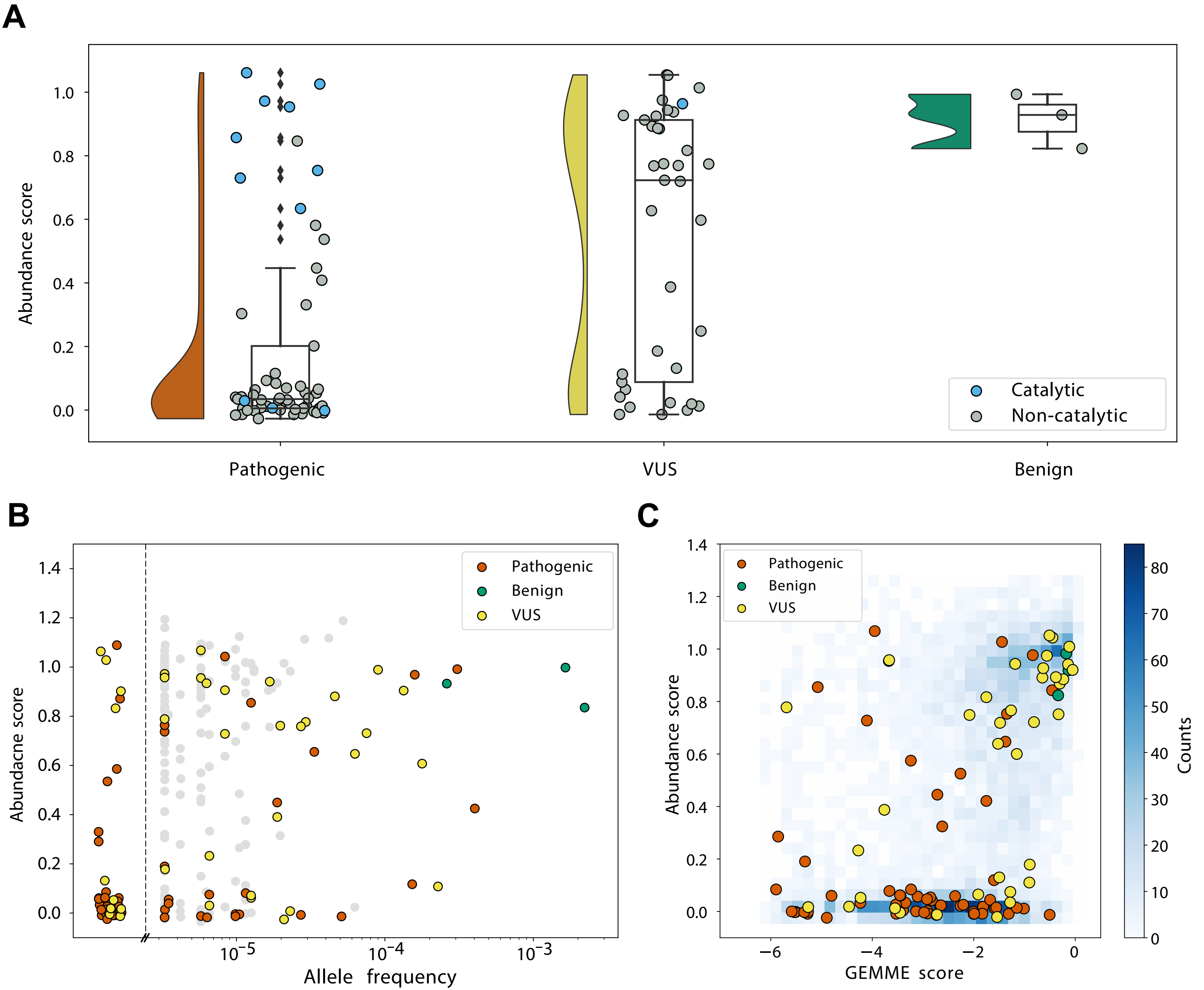

A
Bxb1
barcode
GFP
ASPA
IRES
mCherry
vector DNA
2A
2A
BFP
iCasp9
Blast^R
landing pad
tet ON
Bxb1
recombination
GFP
ASPA
IRES
mCherry
gDNA
tet ON
AAAAAA
mRNA
ASPA
GFP
PQC
mCherry
Proteins
FACS binning
GFP:mCherry ratio
Sorted cells
barcode sequencing
Variant | Abundance score
1 | 1.02
2 | 0.76
... | ...
6152 | 0.09
Variant scoring
B
BFP
mCherry
GFP
merge
WT
C152W
20µm
C
GFP-ASPA
ASPA-GFP
mock
WT
C152W
WT
C152W
70 kDa -
35 kDa -
40 kDa -
GFP (ASPA)
mCherry
GAPDH
D
GFP
10^3
10^2
10^1
10^2
10^3
10^4
mCherry
ASPA library
GFP-WT
GFP-C152W
E
Count
12000
9000
6000
3000
0
10^-2
10^-1
10^0
GFP:mCherry ratio
ASPA library
GFP-WT
GFP-C152W
F
Count
15000
10000
5000
0
1
2
3
4
10^-2
10^-1
10^0
GFP:mCherry ratio
ASPA library

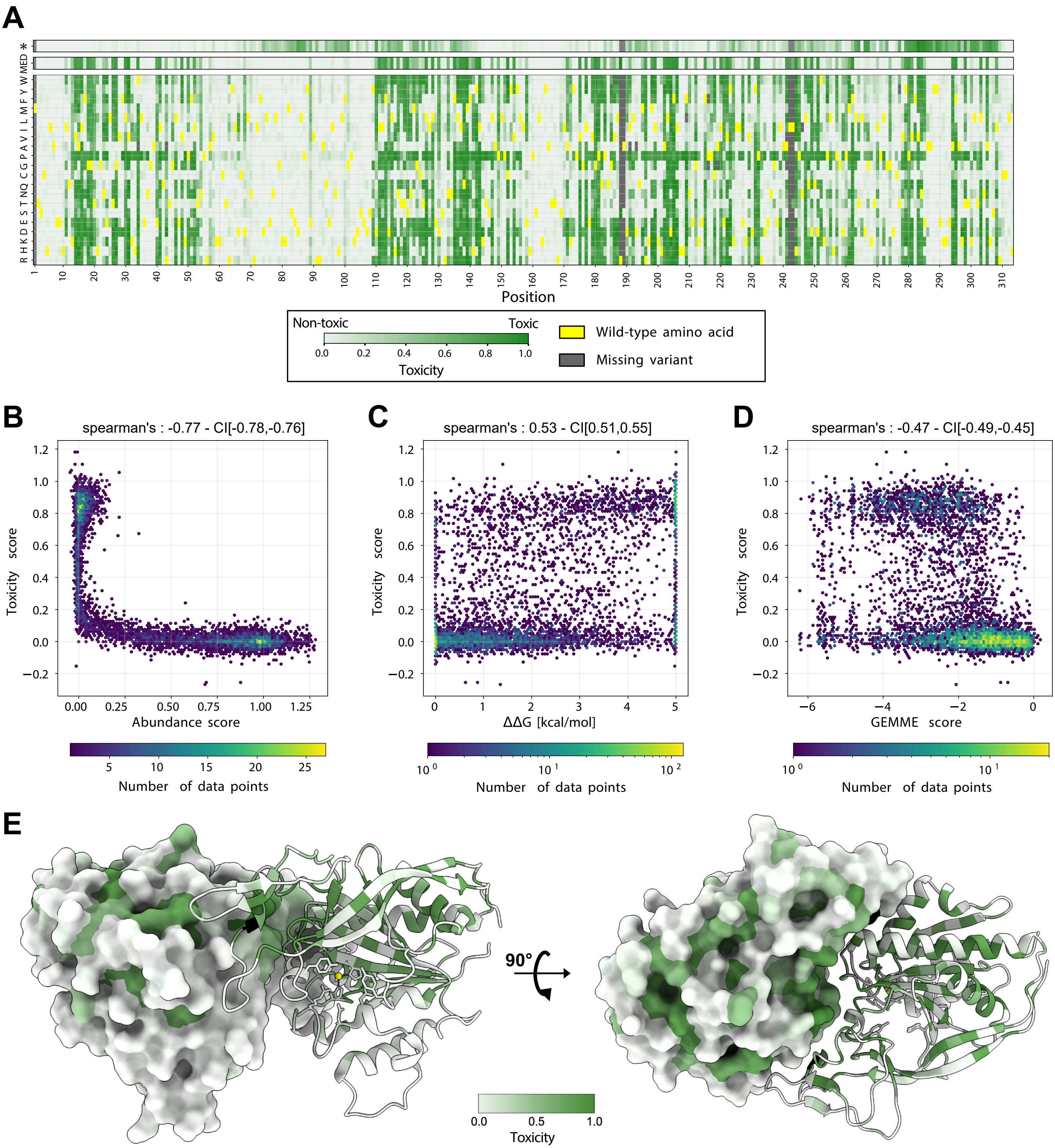

A
*
MED
R H K D E S T N Q C G P A V I L M F Y W
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
Position
pLDDT
ASPA
Disorder region
100
70
50
0
pLDDT
Nonsense
Synonymous
Hyperstable
0.00
0.50
1.00
1.50
Abundance score
Wild−type amino acid
Missing variant
B
C152W
WT
Missense
Synonimous
Nonsense
Counts
0
200
400
600
800
1000
1200
0.0
0.5
1.0
1.5
Abundance score
C
Low throughput abundance score
0.0
0.5
1.0
1.5
L312D
F29H
E285A
M82T
E235K
WT
K163L
L263S
A219L
D114Y
P260E
N121Q
N54K
P257T
T26V
A305E
P254F
C152W
0.0
0.5
1.0
1.5
Abundance score
D
180°


A
Rosetta ΔΔG (thermodynamic stability)
Spearman's ρ : -0.47 - CI[-0.45,-0.49]
Spearman's ρ: - 0.52 - CI[-0.43,-0.60]
Abundance score
ΔΔG [kcal/mol] score
Median abundance score
Median ΔΔG [kcal/mol] score
B
GEMME ΔE
Spearman's ρ: 0.48 - CI[0.46,0.49]
Spearman's ρ: 0.49 - CI[0.40,0.57]
Abundance score
GEMME ΔE
Median abundance score
Median GEMME ΔE
1.2
1.0
0.8
0.6
0.4
0.2
0.0
5
4
3
2
1
0
−6
−5
−4
−3
−2
−1
0


A
ASPA library
GFP-WT
GFP-C152W
8k
6k
4k
2k
0
count
10^{-2}
10^{-1}
10^{0}
GFP:mCherry
B
37 °C
39.5 °C
2.5k
2k
1.5k
1k
0.5k
0
count
10^{-2}
10^{-1}
10^{0}
GFP:mCherry
C
37 °C
29 °C
2.5k
2k
1.5k
1k
0.5k
0
count
10^{-2}
10^{-1}
10^{0}
GFP:mCherry
D
control
+MLN7243
+MLN7243
4k
3k
2k
1k
0
count
10^{-1}
10^{0}
GFP:mCherry
E
control
+BZ
2.5k
2k
1.5k
1k
0.5k
0
count
10^{-2}
10^{-1}
10^{0}
GFP:mCherry
F
control
+CQ
2.5k
2k
1.5k
1k
0.5k
0
count
10^{-2}
10^{-1}
10^{0}
GFP:mCherry
G
control
+YM01
8k
6k
4k
2k
0
count
10^{-1}
10^{0}
GFP:mCherry
H
control
+NAA
8k
6k
4k
1k
0
count
10^{-1}
10^{0}
GFP:mCherry


A
ASPA
ASPA
1
313
24 aa tiles
B
Bxb1
ASPA tile
GFP
IRES
mCherry
vector DNA
2A
2A
BFP
iCasp9
Blast^R
landing pad
tet ON
Bxb1
recombination
GFP
IRES
mCherry
gDNA
tet ON
AAAAAA
mRNA
ASPA
tile
GFP
PQC
mCherry
Proteins
FACS binning
GFP:mCherry ratio
Sorted cells
tile sequencing
ASPA
Tile <WCN>
low
high
C
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
1
2
3
4
TSI
D
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
0.0
0.2
0.4
0.6
0.8
1.0
degron probability


A
Abundance score
1.0
0.8
0.6
0.4
0.2
0.0
Catalytic
Non-catalytic
Pathogenic
VUS
Benign
B
Abundacne score
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Pathogenic
Benign
VUS
10^{-5}
10^{-4}
10^{-3}
Allele frequency
C
Abundance score
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Pathogenic
Benign
VUS
−6
−4
−2
0
GEMME score
Counts
80
70
60
50
40
30
20
10
0


A
*
MED
R H K D E S T N Q C G P A V I L M F Y W
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
Position
Non-toxic
Toxic
0.0
0.2
0.4
0.6
0.8
1.0
Toxicity
Wild-type amino acid
Missing variant
B
spearman's : -0.77 - CI[-0.78,-0.76]
Toxicity score
Abundance score
0.00
0.25
0.50
0.75
1.00
1.25
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5
10
15
20
25
Number of data points
C
spearman's : 0.53 - CI[0.51,0.55]
Toxicity score
ΔΔG [kcal/mol]
0
1
2
3
4
5
10 0
10 1
10 2
Number of data points
D
spearman's : -0.47 - CI[-0.49,-0.45]
Toxicity score
GEMME score
−6
−4
−2
0
10 0
10 1
Number of data points
E
90°
0.0
0.5
1.0
Toxicity