Deep indel mutagenesis reveals the impact of insertions and deletions on protein stability and function

Magdalena Topolska[1,2], Antoni Beltran[1] and Ben Lehner[1,2,3,4]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
2 University Pompeu Fabra (UPF), Barcelona, Spain
3 Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain
4 Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

*e-mail: bl11@sanger.ac.uk

Abstract

Amino acid insertions and deletions (indels) are an abundant class of genetic variants. However, compared to substitutions, the effects of indels are not well understood and poorly predicted. Here we address this shortcoming by performing deep indel mutagenesis (DIM) of structurally diverse proteins. Indel tolerance is strikingly different to substitution tolerance and varies extensively both between different proteins and within different regions of the same protein. Although state of the art variant effect predictors perform poorly on indels, we show that both experimentally-measured and computationally-predicted substitution scores can be repurposed as good indel variant effect predictors by incorporating information on protein secondary structures. Quantifying the effects of indels on protein-protein interactions reveals that insertions can be an important class of gain-of-function variants. Our results provide an overview of the impact of indels on proteins and a method to predict their effects genome-wide.

Introduction

Short insertions and deletions (indels) of amino acids are an abundant source of genetic variation[1–4]. Indels are more challenging to accurately genotype than substitutions[5] and their effects have been much less comprehensively experimentally evaluated[6–11] making it difficult to evaluate how well their impact is computationally predicted. Indeed, many of the most popular variant effect prediction (VEP) algorithms do not even return scores for indels[10].

Indels are a fundamentally different type of perturbation to substitutions. Whereas substitutions are ‘side chain mutations’, indels are ‘backbone mutations’ that can have a much more severe effect on structure, stability and function[12–15]. The most frequent indels in natural genomes are change in copy count (CCC) variants[16] where one or more nucleotides are repeated or deleted due to polymerase slippage during DNA replication[17,18].

Massively parallel DNA synthesis-selection-sequencing experiments - also called deep mutational scanning (DMS) or multiplex assays of variant effect effects (MAVEs) - provide a great opportunity to quantify the effects of indels on proteins at scale in order to better understand their effects and to evaluate and develop computational methods for indel variant effect prediction[6–11]. Here we apply this approach to quantify the impact of diverse indels on the stability of structurally diverse proteins.

We find that deletions and insertions are more detrimental than substitutions and that the site of insertion is more important than the aa inserted. Whereas substitutions are better tolerated on the surface of proteins and most detrimental in their hydrophobic cores, this is less true for insertions or deletions. In general, indels are better tolerated in protein termini than in secondary structure elements, but tolerance varies extensively both between different proteins and within different regions of the same protein. We find that widely used variant effect predictors do not perform well for predicting indel tolerance, but we also show that both experimentally-measured and computationally-predicted substitution scores can be augmented to provide very good genome-scale prediction of aa indel variant effects. Finally, by quantifying protein binding at scale, we find that insertions more frequently generate gain-of-function molecular phenotypes than deletions or substitutions.

Results

Deep Indel Mutagenesis of *in vivo* protein abundance

To better understand and predict the impact of insertions and deletions on proteins, we performed deep indel mutagenesis (DIM) to quantify the impact of 5,624 indels on the cellular abundance of nine structurally diverse protein domains. For each protein, we designed a library of variants covering all sequential deletions of 1 to 3 aa, all change in copy count (CCC) insertions that repeat 1-3 aa, and all 3 nt out-of-frame deletions that remove an aa and substitute an aa in a single event (delSubs) (Fig. 1a). We quantified the effect of each variant on the cellular abundance of each protein using a highly-validated protein fragment complementation assay (PCA), abundancePCA (aPCA), that quantifies protein abundance over three orders of magnitude[19–21]. All transformation-selection-sequencing experiments were performed in triplicate and were highly reproducible

(Extended Data Fig. 1a, Pearson's r=0.920-0.927). aPCA substitution, deletion and insertion scores were also highly correlated with independent measurements of protein stability (Fig. 1b, Extended Data Fig. 1b)[11,22].

Patterns of indel tolerance vary between and within proteins

Our systematic data reveals that all nine proteins are highly intolerant of aa deletions, with 68.8% of 1aa deletions strongly reducing abundance (variants within 1 standard deviation of the deleterious mode, Extended Data Fig. 1c). In general, 1aa insertions are slightly better tolerated, with 62.1% strongly reducing abundance (Extended Data Fig. 1c). However, insertion tolerance varies across proteins: insertions are much better tolerated than deletions in six of the nine proteins, whereas in FBP11-FF1 they are only slightly better tolerated and they are very detrimental in two proteins, GRB2-SH3 and CKS1 (Fig. 1b).

The relative tolerance to insertion and deletion also varies within each protein, with regions similarly tolerant to insertion and deletion, regions more tolerant to deletion, and regions more tolerant to insertion (Fig. 1b). For example, insertions and deletions are similarly tolerated in helix 2 of BL17-NTL9 but deletions are much less detrimental than insertions in helix 1 of the same protein. In CSPA-CSD, insertions and deletions are similarly well-tolerated in loop 3, whereas in loop 1 insertions are more detrimental than deletions, and in loops 2 and 4 deletions are more detrimental than insertions (Fig. 1b).

In general, multi-aa deletions are more detrimental than single-aa deletions (Fig.1c), whereas multi-aa insertions are not more detrimental than single-aa insertions (Fig. 1c). However, again this varies quantitatively across proteins as well as in different regions of the same protein (Fig. 1b). For example, in CI2A-PIN1 multi-aa indels are tolerated similarly to single indels across the entire domain, while in BL17-NTL9 longer indels are more detrimental in certain regions (Fig. 1b). In contrast, in CSPB-CSD double insertions from the middle of loop 1 to the beginning of strand 2 are less detrimental than single and triple insertions in the same region whereas in the rest of the protein single and multi-aa indels have similar effects, with insertions of all lengths being well-tolerated and deletions of all lengths being deleterious (Fig. 1b).

Considering the entire dataset, single aa insertions predict the effects of single aa deletions reasonably well (Pearson's r=0.473, Extended Data Fig. 1d), but the effects of double and triple aa insertions are less well correlated with the effects of double and triple aa deletions (double: r=0.268; triple: r=0.146; Extended Data Fig. 1e and 1f)). Indeed, while single insertions predict the tolerance to longer insertions (r=0.718 and r=0.726 for 2 and 3 aa insertions; Extended Data Fig. 1g and 1h), double and triple deletion tolerance is less correlated with single deletion tolerance at the same position (r=0.617 and r=0.409 for 2 and 3 aa deletions; Extended Data Fig. 1i and 1j).

Thus, in general, deletions are more detrimental than insertions, multi-aa deletions are more detrimental than 1 aa deletions, multi-aa insertions are tolerated similarly to 1aa insertions, and deletion and insertion tolerance are related. However, these relationships vary between different proteins and between different regions within the same protein.

Comparing the impact of insertions, deletions and substitutions

For two of the proteins, PSD95-PDZ3 and GRB2-SH3, in addition to the CCC insertion repeats, we also quantified the effects of all 20 aa insertions at every position. For the same proteins we also measured the effects of all 19 aa substitutions at every site, allowing us to directly compare the effects of inserting and substituting to the same residues.

For PSD95-PDZ3, which has a typical moderate insertion tolerance (Fig. 2a), the average effect of substitutions predicts reasonably well the effect of deleting the residue (r=0.534, Fig. 2b) as well as CCC insertions after (r=0.441, Fig. 2b) and before (r=0.313, Extended Data Fig. 2a) the site. For GRB2-SH3, which is unusually intolerant of insertions (Fig. 2a), substitution tolerance is reasonably predictive of CCC insertions after the substitution site (r=0.392) but much less predictive of the tolerance for insertions before a site (r=0.057 [p=0.693], Extended Data Fig. 2a) and deletion tolerance (r=0.194 [p=0.173], Fig. 2b). Across 175 protein domains with *in vitro* fold stability measurements of indels and substitutions[11], substitution tolerance predicts Ala insertion tolerance similarly well after (mode r=0.281) and before a residue (r=0.240) and better predicts 1 aa deletion tolerance (mode r=0.459, Fig. 2c and Extended Data Fig. 2b).

Interestingly, substitution tolerance is a better predictor of insertion and deletion tolerance in loops than in secondary structure elements (mode r=0.576 and r=0.546 for insertions and deletions in loops, mode r=0.082 and r=0.326 for insertions and deletions in secondary structure elements, Fig. 2c). That substitution tolerance predicts indel tolerance reasonably well - and variably so in different protein regions - suggests a strategy for predicting indel variant effects (see below).

Insertions of different amino acids

Although CCC insertions are by far the most frequent aa insertions in natural genomes[16], experimentally we can insert any aa at any position. Substitutions to different aa often have strikingly different effects at the same residue (Fig. 2a and Extended Data Fig. 2d). Such variation is also seen for insertions, although it is quantitatively less important, with a lower standard deviation of variant effects at most positions (Fig. 2d). However, this again varies across sites, with insertions to different aa having very different effects at some positions (Fig. 2e). For example, at positions 4 and 15 in GRB2-SH3 and positions 14 and 20 in PSD95-PDZ3 insertions have a wide range of effects. At other sites most insertions are tolerated and only particular aa are detrimental. Examples include positions 17 and 53 in GRB2-SH3 and positions 3 and 39 in PSD95-PDZ3 (Fig. 2e).

Considering all sites, there is a reasonable correlation between the average effect across all residues of inserting an aa and substituting to the same aa in both proteins (r=0.530 and r=0.598 for GRB2-SH3 and PSD95-PDZ3, respectively, Fig. 2f). In GRB2-SH3, for example, insertions to Lys, Gly and Cys are highly tolerated as are substitutions to Lys and Cys but not substitutions to Gly (Fig. 2f). However, the tolerance to insertion of different aa varies between the two proteins, between secondary structure elements and loops within the same protein, and between residues within each structural element (Fig. 2e, Extended Data Fig. 2e).

DelSubs

Out of frame 3nt deletions can cause a more complicated protein sequence change where one aa is deleted and the next aa is substituted. We refer to these variants as ‘delSubs’. An important example of a pathogenic delSub is the F508del variant in the cystic fibrosis transmembrane conductance regulator (CFTR) gene, which is the most frequent cause of cystic fibrosis[23] .

To better understand delSubs, we quantified the effects of 325 across the nine structurally diverse proteins. Our data shows that the effects of delSubs correlate well with the effects of deletions of the same residues (r=0.729, Fig. 2g), showing that delSub tolerance is predominantly driven by deletion tolerance. However, for a subset of sites where deletions are tolerated, delSubs are detrimental. At these sites, which constitute ~8% of delSubs, the additional substitution results in destabilisation of the protein (Fig. 2g).

Structural determinants of indel tolerance

The quantification of thousands of insertion, deletion and substitution variants across diverse protein folds provides an opportunity to better understand how their effects relate to structure. Consistent with many previous experimental[7,10] and computational analyses[15,24] , we find that substitutions in the hydrophobic cores of proteins are much more detrimental than substitutions in solvent exposed surface residues (correlation with relative solvent-accessible surface area, rSASA, r=0.645, r=0.756 for GRB2-SH3, PSD95-PDZ3 and mode r=0.682 for *in vitro* fold stabilities of 178 domains, Fig. 3a-b). However, solvent accessibility is a less good predictor of deletion (r=0.151 [p=0.290], r=0.487, mode r=0.388) and insertion (r=0.0246 [p=0.863], r=0.288, mode r=0.144) tolerance (Fig. 3b). Patterns of substitution and indel tolerance also differ with respect to secondary structure and location in a protein. In alpha-helices, substitution tolerance follows the 3-4 aa structural periodicity of helices, consistent with substitutions of side chains on the same face of a helix having more similar effects (Fig. 3c). Similarly, in beta-strands, substitutions in every second aa have more similar effects, consistent with their side chains being on the same face of a strand (Fig. 3d). Neither of these periodicities in variant effects is observed for insertions or deletions (Fig. 3c/d, lower panels). Moreover, insertions and deletions are well tolerated in the N- and C-termini of proteins (termini defined as the sequence before or after the first or last secondary structure element, respectively) but are more detrimental than substitutions in loops, helices, and strands (Fig. 3e). Indel tolerance also varies more than substitution tolerance depending upon the length of secondary structure elements and the identity of the structural element before and after that in which the mutation occurs (Extended Data Fig. 3c-e). In protein termini, insertions and deletions are more detrimental close to the start or end of secondary structures (Extended Data Fig. 3b).

In summary, unlike substitution tolerance, indel tolerance does not relate strongly to the periodicity of secondary structure elements and the solvent exposure of side chains. Rather, multiple features seem potentially important, including: i) the secondary structure element in which the indel is located, ii) the length of that element, iii) the neighbouring environment of the element, and iv) the exact residue. Some or all of these features may therefore be useful variables for indel variant effect prediction.

Evaluating indel variant effect prediction

We next evaluated how well computational variant effect prediction (VEP) methods predict the effects of indels. We evaluated the performance of two widely used VEPs, CADD[25] and PROVEAN[26], as many other widely-used methods do not return scores for indels (Extended Data Table 1). CADD only provides predictions for human proteins and its performance is reasonable for substitutions (mode r=0.389, Fig. 4a) but poor for insertions (mode r=-0.068) and deletions (mode r=-0.026, Fig. 4a) In contrast, PROVEAN performs better for insertions (mode r=0.440) and deletions (mode r=0.299) than for substitutions (r=0.267, Fig. 4a), but the performance varies extensively across proteins (interquartile range from r=0.250 to r=0.582 for insertions and r=0.259 to 0.570 for deletions). Interestingly CADD and PROVEAN predictions correlate well for substitutions (r=0.402, Extended Data Fig. 4a) but not for insertions (r=0.0831) and deletions (r=0.164, Extended Data Fig 4b).

Accurate indel variant effect prediction

We next explored whether a range of relatively simple models could be used to predict the effects of indels on fold stability. The experimentally measured average substitution scores provide reasonably good prediction of the effects of deletions (model 1, mode r=0.464 for 1aa deletions, evaluated by leave-one-protein-out cross validation, Fig. 4c) but less good prediction for insertions (mode r=0.240 for Ala insertions) (Fig. 4d). The identity of the aa at each residue is not a good predictor of indel tolerance (model 2, Fig. 4 c/d). A regression model using only secondary structure information (model 3) is quite a good predictor of both deletion (mode r=0.550) and insertion (mode r=0.561) tolerance (Fig. 4 c/d). The secondary structure features used in model 3 include the secondary structure element in which the indel occurs, its length, the position of the indel within the element and the identity of the neighbouring secondary structure elements (see Methods). Better performance for both insertions and deletions is obtained when using secondary structure information and the mean substitution score per position (model 5, mode r=0.617 and r=0.613, for deletions and insertions; adding information about the starting aa did not improve performance, model 4, Fig. 4 c/d).

Repurposing substitution variant effect predictors for indels

We next explored whether we could replace the experimentally-measured substitution scores with computationally-predicted scores. State-of-the-art deep learning methods are showing increasingly good performance for predicting substitution variant effects[27–32] . We evaluated two methods: ESM1v, an unsupervised large language model[29] , and DDMut, a deep neural network trained specifically to predict stability changes[32]. Not unexpectedly, we found that DDMut provides better prediction for the effects of substitutions on stability changes (mode r=0.567 for ESM1v, r=0.664 for DDmut across 181 domains, Fig. 4a). We therefore tested how well the DDMut substitution scores could replace the experimental substitution scores in our regression models. Predictive performance using the DDmut substitution scores alone is lower than when using the experimental substitution scores (model 1p, r=0.259 and r=0.156 for deletions and insertions, Fig. 4c/d). However, combining these scores with secondary structure information results in good predictive performance (model 5p, median r=0.613 and r=0.616), similar to the performance when using experimental substitution scores (model 5, median r=0.617 and r=0.613).

The coefficients of the best performing models (models 5 and 5p) highlight the experimental and computational substitution scores as key predictors of destabilisation (Extended Data Fig. 4c and 4d). Additional features include indels within or in structural elements connected to alpha helices and strands and indels in termini residues adjacent to secondary structures. Features predicting stabilisation involve indels at the protein's ends and in loops or termini, particularly those distant from the initial or final secondary structure element.

We conclude that the combination of either experimentally-measured or computationally-predicted substitution scores with secondary structure information provides good prediction of the effects of indels across 178 different proteins.

Insertions generate gain-of-function molecular phenotypes

Destabilisation is likely to be the most frequent mechanism by which missense variants cause disease[33–37]. However, proteins normally have diverse molecular activities beyond folding that can also be affected by mutation and disease mechanisms. Moreover, an important subset of disease variants have gain-of-function mechanisms[38]. We therefore also quantified the effects of substitutions, insertions and deletions on a biophysical activity beyond fold stability, focussing on protein binding, which is a molecular function important for nearly all proteins. Using a highly-validated quantitative protein-protein interaction selection assay[20,21,39], Fig. 5a), we quantified the effects of all aa substitutions and diverse insertions and deletions on the binding of GRB2-SH3 and PSD95-PDZ3 to their respective ligands, GRB2-associated binding protein 2 (GAB2) and CRIPT (Fig. 5c).

In both proteins, many insertions and deletions strongly reduce binding to their ligands, consistent with their effects on fold stability (Fig. 5b,d,e). Most strikingly, however, a large number of 1 aa insertions in PSD95-PDZ3 have gain-of-function phenotypes, strongly increasing its binding to the CRIPT ligand (Fig. 5c,d). These insertions are nearly all in loops 1 and 2 of the protein (Fig. 5c,f). A subset of multi-aa insertions also increase binding (Extended Data Fig. 5b). Insertions in these loops have a much stronger gain-of-function molecular phenotypes than substitutions (Fig. 5c). Loop 1 constitutes the carboxylate-binding loop, directly involved in peptide recognition and forming the binding site groove for CRIPT together with strand 2 and helix 2[41]. The second loop connects strand 2 and 3, which both contain ligand-contacting residues[41]. The insertions in these loops that increase binding have only a mild effect on the abundance of PSD95-PDZ3 (Fig. 2a)

Thus, although insertions and deletions - like substitutions - primarily have loss-of-function molecular phenotypes, insertions in two loops of PSD95-PDZ3 cause strong gain-of-function phenotypes.

Discussion

Here we have used deep indel mutagenesis to quantify, understand and learn how to predict the effects of insertions and deletions in proteins. Indels and substitutions are fundamentally different changes to proteins, and this is reflected in their different patterns of tolerance. Our data show that tolerance varies both between different proteins and different regions of the same protein, as does the relative tolerance to insertion and deletion.

Despite this apparent complexity, we have also shown that it is possible to predict the effects on indels using relatively simple models that combine experimentally-measured or computationally-predicted substitution scores with information about the secondary structure of a protein. State-of-the-art substitution variant effect predictors can thus be augmented and repurposed as good indel predictors, which is important given the larger training and evaluation datasets currently available and being produced for substitutions[43,44]. These models are deliberately simple and interpretable, and it is likely that machine learning approaches will further improve performance.

Finally, we have presented evidence that insertions can be an important source of gain-of-function molecular phenotypes, which are particularly challenging variants to identify and predict across diseases[38].

The expansion of deep indel mutagenesis to additional proteins and to additional molecular and cellular phenotypes is an important goal for future work. Large experimental indel mutagenesis will provide training and evaluation data for computational models for different protein properties and reference atlases to guide the interpretation of clinical variants. Large-scale indel mutagenesis and models to predict the effects of indels are, moreover, likely to be important for protein engineering, providing access to gain-of-function and change-of-function phenotypes that are difficult to access through substitutions alone.

Materials and methods

Deep Indel mutagenesis library design

We designed the mutational libraries using sequences listed in Extended Data Table 2 as the wildtype templates. For GRB2-SH3 and PSD95-PDZ3 we designed i) 1-3 aa neighbouring deletions ii) 1-3 aa neighbouring CCC insertions iii) out-of-frame 3nt deletions resulting in delSubs iv) 15x synonymous mutations prioritising >1nt changes v) all possible 19 substitutions at each position using most abundant yeast-codons (https://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932) and prioritising >1nt changes in the substituted codons vi) all possible 20 aa insertions at each position vii) 1-3 aa repeats of each aa in the template sequence viii) 2aa insertions of Cys and viv) 2x insertions of Ala. For the remaining 7 protein domains, we included i-iv into the mutational design. The 5' and 3' prime adapters were added to each protein for amplification and cloning purposes and adjusted to each wild type (wt) template length, with final adapters ranging from 20-42nt. The library was ordered at Twist Biosciences as a pooled oligo library with final lengths of single stranded oligos ranging from 149-250nt. The Twist pool was resuspended in water to the concentration of 1 ng/uL and the 3 sub pools 1) GRB2-SH3 2) PSD95-PDZ3 and 3) 7 domains were separated by PCR amplification of 14 cycles, using primers listed in Extended Data Table 3 and using 10ng of the Twist pool as template in each reaction. The 3 library pools were purified by an ExoSAPII (NEB) reaction to remove single-stranded DNA and further by column purification (MiniElute Gel Extraction Kit, QIAGEN).

Variant library construction

We used generic abundance- and bindingPCA plasmids[20,21] to construct the 3 sub libraries. GRB2-SH3 and PSD95-PDZ3 libraries were cloned into their respective aPCA plasmids pGJJ046 and pGJJ068 (N-terminal DHFR tag[20]) using Gibson assembly. The 7 domain library was cloned into the pGJJ162 plasmid (C-terminal DHFR tag[21]). The backbones for the Gibson reaction for GRB2-SH3 and PSD95-PDZ3 library assembly (aPCA plasmids) were first linearized using primers listed in Extended Data Table 3 and next treated with Dpn1 (NEB) restriction enzyme to remove the circular plasmid template. Next the correct sizes of the linearized backbones were isolated using gel electrophoresis and later purified using QIAquick Gel Extraction Kit (QIAGEN). For the library of 7 domains, the pGJJ162 backbone was linearised by digestion with restriction enzymes NheI and HindIII (NEB) for the Gibson reaction. We used 100ng of the linearized vectors and 9.4ng, 9.28ng and 11.23ng of the purified double-stranded libraries (GRB2-SH3, PSD95-PDZ3 and 7 domains) for each gibson reaction of 20 uL. The gibson reaction (in-house prepared enzyme mix) was incubated at 50C for one hour, then desalted by dialysis using membrane filters for 1 hour, concentrated to 3X by SpeedVac (Thermo Scientific) and transformed into NEB 10β High-efficiency Electrocompetent E. coli cells according to the manufacturer's protocol. Cells were allowed to recover in SOC medium (NEB 10β Stable Outgrowth Medium) for 1 hour and later transferred to 100 mL of LB medium with ampicillin overnight. 100 mL of each saturated E. coli culture were harvested next morning to extract the plasmid library using the QIAfilter Plasmid Midi Kit (QIAGEN). Finally, after verifying correct assembly sanger sequencing (Eurofins), the GRB2-SH3 and PSD95-PDZ3 libraries were digested out of the aPCA plasmids using NheI and HindIII for re-assembly of the bPCA library. The bPCA library was assembled overnight by temperature-cycle ligation using T4 ligase (New England Biolabs) according to the manufacturer's protocol, 67 fmol of backbone and 200 fmol of insert in a

33.3 uL reaction. As backbones for GRB2-SH3 and PSD95-PDZ3 library inserts, we used the pGJJ034 and pGJJ072 plasmids with GAB2 and CRIPT ligands fused N-terminally to DHFR12, as described in our previous study[20]. The ligation was desalted by dialysis, concentrated 3X, transformed into NEB 10β High-efficiency Electrocompetent E. coli cells, and purified from E. coli using the QIAfilter Plasmid Midi Kit as described above.The coverage of variants after each transformation reaction was estimated to be >20x.

AbundancePCA and BindingPCA selections

aPCA and bPCA methotrexate selection were performed as described in our previous studies[20,21]. The high-efficiency yeast transformation protocol was scaled to 100 mL based on the targeted number of transformants of each library and each biological replicate was transformed into cells grown from independent colonies of BY4741 yeast strain (https://www.yeastgenome.org/strain/by4741). In total we completed 5 methotrexate selection assays of 3 replicates each, 3 aPCA for GRB2-SH3, PSD95-PDZ3 and the 7 domain library, and 2 bPCA for GRB2-SH3 and PSD95-PDZ3 libraries. The pre-selection medium used was SC-URA/ADE and the selection medium was SC-URA/ADE + 200ug/mL Methotrexate (BioShop Canada Inc., Canada). Harvested input and output cells were pelleted, washed with water and stored at -20C until the DNA extraction step.

DNA extractions and plasmid quantification

The DNA extraction protocol was used as described in our previous study[20,21]. We extracted DNA from a 50mL harvested selection input and output cultures at OD600nm~1.6. Plasmid concentrations in the total DNA extract (that also contained yeast genomic DNA) were quantified by qPCR using the oligo pair 6 (Extended Data Table 3), that binds to the ori region of the plasmids.

Sequencing library preparation

Sequencing library preparation was done as described in our previous study[20,21].We performed 2 consecutive PCR reactions for each sample. The first PCR (PCR1) is used to amplify the amplicons for sequencing, to add a part of the illumina sequencing adaptors to the amplicon and to increase nucleotide complexity for the sequencing reaction by introducing frame-shift bases between the adapters and the sequencing region of interest. PCR1 frame-shifting (fs) oligos for each of the sub libraries are listed in Extended Data Table 3. The second PCR (PCR2) is used to add the remainder of the illumina adaptor and to add demulitplexing indexes. All samples, except the GRB2-SH3 bPCA, were dual-indexed using differing barcode indexes both for the forward (5' P5 Illumina adapter) and reverse oligos (3' P7 Illumina adapter). The GRB2-SH3 bPCA library was single-indexed using a constant forward oligo (3' P7 Illumina adapter) and alternating reverse oligos (3' P7 Illumina adapter). The demulitplexing primers used for PCR2 are listed in Extended Data Table 4. The amplicon library pools were isolated based on size by gel electrophoresis using a 2% agarose gel and then purified using QIAEX II Gel Extraction Kit (QIAGEN) and using 30uL of QIAEX II beads for each sample. The purified amplicons were subjected to Illumina 150bp paired-end NextSeq 2000 sequencing at the CRG Genomics Core Facility.

Sequence data processing

FastQ files from paired-end sequencing of all aPCA and bPCA experiments were processed with DiMSum[45] v1.2.11 using default settings with minor adjustments: https://github.com/lehner-lab/DiMSum. Due to low read coverage, following samples were re-sequenced: GRB2-SH3 bPCA input replicate 3, output replicate 1 and 2. Outputs from the second run of sequencing were added as technical replicates in the Experimental Design File when running DimSum for the GRB2-SH3 bPCA experiment. Variant counts associated with all samples (output from DiMSum stage 4) were filtered using a custom script to retain only the programmed variants (Variant Identity File). FastQ files were processed with DiMSum separately for the GRB2-SH3 and PSD95-PDZ3 aPCA and bPCA samples and in bulk for the remaining 7 domains. For the 7 domains, we also supplied the Synonym Sequence File listing all wt sequences, to retrieve the synonymous variants for all 7 domains.

The Dimsum output "...fitness_replicates.RData" files containing the fitness and fitness error estimates were used for further data analysis. After calling the substitution, indel and synonymous variants for each protein, we next normalised the data for each protein separately. To determine the lower limit for normalisation, we applied the Chernoff mode estimator (using the mlv{modeest} function) to identify the mode of the lower peak in the bimodal distribution of all variant effects.This value was subtracted from the DimSum "fitness" value resulting in "norm_fitness". The "norm_fitness" of each variant was then divided by the weighted mean of the synonymous variants, resulting in "scaled_fitness". For one of the proteins, VIL1-HP, we used the "Vieu" mode estimator of the mlv{modeest} function, due to non-overlapping deleterious modes of insertions and deletions. We normalised the replicate errors by dividing the DimSum fitness sigma with the square root of the weighted mean of synonymous variants. Indel variants with >6 and 0 output counts were added back into the working dataframe from the "...variant_data_merge.RData" file, as DimSum eliminates everything with 0 counts in the output. These variants were only CKS1 mutants (34 out of 432) and were assigned "scaled_fitness" of 0 as they are thought to be completely deleterious for protein abundance. For all variants, we tested if changes in abundance were significant from the weighted mean of the synonymous variants by calculating z-stats with a two-tailed test and using Bonferroni multiple testing correction. The highly deleterious CKS1 variants with 0 counts in the DiMSum output were not marked as significant changes.

In vitro fold stability data

In vitro fold stability data from Tsuboyama et al.[11] was downloaded from https://zenodo.org/record/7992926. The downloaded files used for further analysis were "Tsuboyama2023_Dataset2_Dataset3_20230416.csv" containing the inferred deltaG and ddG scores for all substitution and indel variants and the "AlphaFold_model_PDBs " folder containing the pdb files for all assayed domains. For our analysis, we inverted the sign of the inferred "ddG_ML" by multiplying it with -1. Additionally, we filtered the original data frame, retaining only the natural protein domains and domains with scores for all mutation types (substitutions, single deletions and alanine insertions) resulting in a final selection of 181 protein domains.

Secondary structure features

Secondary structure assignments were made using STRIDE[46] using the stride{bio3d} R function. The secondary structure alignment was simplified from 6 categories ("AlphaHelix," "310Helix," "Strand," "Turn," "Coil," "Bridge") by merging annotations for "Turn," "Coil," and "Bridge" into a single category called "loop". Furthermore, we annotated the N- and C-termini as the sequence of amino acids prior to or immediately after the first residue of the first/last secondary structure element ("AlphaHelix", "310Helix" and "Strand"). Next, we realigned every secondary structure element using the relative solvent accessibility score calculated using PyMol[47] v.2.3.5. For each structure element ("AlphaHelix", "310Helix", "Strand" and "loop") we first found the median position based on the element length and set the position 0 to the most buried residue (based on the rSASA score) +1/-1 from the median length. Residues towards the n-terminal of the secondary structure element were annotated in negative, descending order (-1, -2, -3 etc) while the residues towards the c-terminal were annotated in ascending order (1, 2, 3 etc). For the N- and C-termini, we annotated the first position immediately prior to or after the secondary structure as position -1/+1. For N-termini we annotated the rest of the positions in negative, descending order while for the C-termini the positions were annotated in an ascending order.

Variant effect prediction

CADD[25] was run on the human subset of the domains from Tsuboyama et al. [11]. The chromosomal coordinates of each aa sequence were manually annotated using the UCSC Genome Browser[48] (https://genome.ucsc.edu/index.html) and the VCF files containing all substitutions and single indels coordinates together with the reference- and alternative sequences submitted through the CADD web interface at https://cadd.gs.washington.edu/score. PROVEAN[26] predictions were run locally using version 1.1.5, available for download at: https://www.jcvi.org/research/provean#downloads. For the PROVEAN input we used the wildtype aa sequences from the Tsuboyama et. al. dataset from which we encoded all possible substitutions and single indels. ESM1v[29] (https://github.com/facebookresearch/esm) was run with minor modifications to the code (predict.py) to allow execution on a CPU with a pytorch installation without CUDA support. DDMut[32] was run using the Application Programming Interface (API) with further instructions available at https://biosig.lab.uq.edu.au/ddmut/api. We encoded all possible substitutions using wildtype aa-sequences from the Tsuboyama et. al. dataset. The pdb files used for the DDMut submission were those provided for the Tsuboyama et al. dataset at https://zenodo.org/record/7992926.

Model design

For the deletion and insertion prediction models we used multiple linear regression with lasso regularisation without interaction terms, which allowed us to fit a simple linear model to the data while encouraging sparsity of the predictive features by shrinking some regression coefficients to zero. The predictive features (dummy-) encoded for the models 1-5 were: "resid", "simple_struc", "structure_before", "structure_after", "align_to_centre", "align_to_centre_termini", "length", "ddG_ML_subs" and "ddMut". The "resid" had 20 levels which described the wildtype aa of the deletion position or the wildtype position before the inserted aa. The 5 levels of secondary structure elements ("AlphaHelix", "Strand", "310Helix", "loop" and "termini") were encoded by "simple_struc". The "structure_before" and "structure_after" encoded the structural element immediately before or after the current element and had 6 levels ("start"/"end", "ntermini"/"ctermini", "loop", "Strand", "AlphaHelix", "310Helix"). The re-aligned position of each secondary structure element was encoded by "align_to_centre" and was simplified to 9 levels describing the position 0, the 1st, 2nd and 3rd position prior to or after 0, while the rest of the residues were labelled as ">4+" or "-4<". The same strategy was applied for simplifying the encoding of termini positions in "align_to_centre_termini". For the "length" we encoded simplified length of each secondary structure element described by 3 levels, "short", "medium" and "long", which was based on the frequency of the the element lengths across the Tsuboyama et al. dataset and adapted to each secondary structure element individually. Finally, we used the mean inferred ddG of substitutions/residue, “ddG_ML_subs”, as a predictive feature for the model 5 and “ddMut”, the mean predicted ddG of stability for 19 substitutions/residue for model 5p. The models were evaluated using leave-one-out cross-validation where the model was trained on all-except-one domain, and evaluated on that held-out domain. For each individual model, using R, we first determined the optimal regularisation parameter (lambda) by cross-validation using the cv.glmnet{glmnet} function with lasso penalty (alpha=1). Here we also calculated the best lambda value, representing the optimal regularisation strength that minimises overfitting while maintaining model performance. The best-fitting model was trained on all-except-one domain using the glmnet{glmnet} function with the selected optimal lambda from the previous step. Finally, the model performance was tested on the held-out domain. Therefore, in conclusion, the deletion and insertion predictors were tested 181 times for each domain. The final figures contain evaluations of the models for 178 domains as we filtered for >3 data points when calculating the correlation coefficients.

Data availability

All DNA sequencing data have been deposited in the Gene Expression Omnibus under the accession number GSE244096. All scaled fitness measurements for i) aPCA, ii) aPCA of selected substitution mutants for the *in vitro* ddG validation, iii) bPCA as well as iv) the processed Tsuboyama et al. data for indels and mean substitutions/residue and v) all single substitutions are available as Extended Data files 1-5 respectively.

Code availability

Source code used to perform all analyses and to reproduce all figures in this work is available at: https://github.com/lehner-lab/deep_indel_mutagenesis.

Author contributions

M.T. performed all experiments and analyses except the generation of aPCA substitution data for comparison with *in vitro* ddG measurements and ESM1v predictions that were performed by T.B.. M.T. and B.L. conceived the project, designed analyses, and wrote the manuscript with input from T.B.

Acknowledgements

This work was funded by a European Research Council (ERC) Advanced (883742) grant, the Spanish Ministry of Science and Innovation (LCF/PR/HR21/52410004, EMBL Partnership, Severo Ochoa Centre of Excellence), the Bettencourt Schueller Foundation, the AXA Research Fund, Agencia de Gestio d’Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. M.T. was funded by a Spanish Ministry of Science and Innovation Severo-Ochoa fellowship PRE2020-093984-SO). T.B was funded by an EMBO (ALTF 183-2020) and Marie Skłodowska-Curie (101030961) fellowship.

References

1. Miton, C. M. & Tokuriki, N. Insertions and Deletions (Indels): A Missing Piece of the Protein Engineering Jigsaw. *Biochemistry* **62**, 148–157 (2023).

2. Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **7**, 9313 (2017).
3. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).
4. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–6 (2010).
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
6. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
7. Gonzalez, C. E., Roberts, P. & Ostermeier, M. Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β-Lactamase. *J. Mol. Biol.* **431**, 2320–2330 (2019).
8. Ogden, P. J., Kelsic, E. D., Sinai, S. & Church, G. M. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**, 1139–1143 (2019).
9. Seuma, M., Lehner, B. & Bolognesi, B. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat. Commun.* **13**, 7084 (2022).
10. Macdonald, C. B. *et al.* DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. *Genome Biol.* **24**, 36 (2023).
11. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
12. Shortle, D. & Sondek, J. The emerging role of insertions and deletions in protein engineering. *Curr. Opin. Biotechnol.* **6**, 387–393 (1995).
13. Arpino, J. A. J., Rizkallah, P. J. & Jones, D. D. Structural and dynamic changes associated with beneficial engineered single-amino-acid deletion mutations in enhanced green fluorescent protein. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2152–2162 (2014).
14. Schenkmayerova, A. *et al.* Engineering the protein dynamics of an ancestral luciferase. *Nat. Commun.* **12**, 3616 (2021).
15. Savino, S., Desmet, T. & Franceus, J. Insertions and deletions in protein evolution and engineering. *Biotechnol. Adv.* **60**, 108010 (2022).
16. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).
17. Streisinger, G. *et al.* Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 77–84 (1966).
18. Garcia-Diaz, M. & Kunkel, T. A. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem. Sci.* **31**, 206–214 (2006).
19. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep.* **7**, 1333–1340 (2014).
20. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
21. Weng, C., Faure, A. J. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *bioRxiv* 2022.12.06.519122 (2022) doi:10.1101/2022.12.06.519122.
22. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021).
23. Collins, F. S. *et al.* Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**, 1046–1049 (1987).
24. Tóth-Petróczy, A. & Tawfik, D. S. Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol. Biol. Evol.* **30**, 761–771 (2013).
25. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human

genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
26. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
27. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
28. Jagota, M. *et al.* Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biol.* **24**, 182 (2023).
29. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.
30. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
31. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
32. Zhou, Y., Pan, Q., Pires, D. E. V., Rodrigues, C. H. M. & Ascher, D. B. DDMut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Res.* **51**, W122–W128 (2023).
33. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).
34. 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).
35. Gao, M., Zhou, H. & Skolnick, J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure* **23**, 1362–1369 (2015).
36. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J. Mol. Evol.* **82**, 11–16 (2016).
37. 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).
38. Backwell, L. & Marsh, J. A. Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–498 (2022).
39. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**, (2018).
40. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59–69 (2003).
41. Doyle, D. A. *et al.* Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076 (1996).
42. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* **7**, (2018).
43. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *Paperpile* https://paperpile.com/app/p/bb90a9c3-64b0-07dd-8db6-44361e1a83f9.
44. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023).
45. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
46. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500–2 (2004).
47. Schrödinger, L. L. C. *The PyMOL Molecular Graphics System*.
48. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Figure legends

Figure 1. Deep indel mutagenesis of diverse protein domains. a. Library design including single and multi-aa indels, delSubs and single aa substitutions. **b**. Deep indel mutagenesis overview for each of the 9 domains. For each domain we show: i) correlation between *in vitro* measured ddG stability and the aPCA scores of selected substitution mutants. Selection of mutants was based on available *in vitro* ddG scores in Protherm[22] ii) density distributions of effects for 1-3 aa CCC insertions (blue), 1-3 aa deletions (green) and single substitutions (purple); black line: weighted mean of the synonymous mutants; red line: mode of the deleterious distribution iii) domain structures coloured by the mean effects of 1-3 aa deletions and CCC insertions iv) heat maps of protein abundance effects of all mutation types with significant changes in abundance (Bonferroni multiple testing correction of z-stats in a two-tailed test) marked with "*" v) effects of 1 aa deletions (green) and 1 aa CCC insertions (blue) across the domain length with the plot background colour-coded by secondary structure elements; white: loop/termini; blue: strand; red: helix (lower panel). **c**. Variation in indel deleteriousness across 1-3 aa indel lengths. P-values were calculated using Wilcox two-sided t-test with multiple testing adjustment (Bonferroni).

Figure 2. Substitutions and insertions to different aa. a. Heatmaps of 19 single aa substitutions, 20 aa insertions, 1-3 aa insertion repeats and 1-3 aa deletion effects on protein abundance. Changes in abundance significant from the weighted mean of the synonymous variants are marked with "*" (Bonferroni multiple testing correction of z-stats in a two-tailed test); y-axis: identity, type or length of mutation; x-axis: mutated position- and the aa sequence of the domain. The x-axis is coloured by a secondary structure element; black: termini/loop; red: helix; blue: strand residues. **b**. Scatter plots of 1 aa deletion or 1 aa CCC insertion versus effects of mean substitution/residue with Pearson's correlations. **c**. Correlations of indel and substitution effects for 175 domains with >3 scores from Tsuboyama et al., and correlation between mutational types within secondary structure regions (black) and loops (white). **d**. Variation of substitution (purple) and insertion (blue) effects per position as standard deviation of effects/residue. **e**. Protein abundance scores of 20 aa insertions at each position; blue dot: insertions of alanine **f.** Scatter plots of mean insertion and mean substitution scores/aa identity with Pearson's correlations. **g**. Scatter plot of 1 aa deletions and 1 aa delSub abundance scores and the Pearson's correlation coefficient.

Figure 3. Structural determinants of indel and substitution tolerance. a. Scatter plots of protein abundance scores for mean substitution/residue (n=52), 1 aa CCC insertion and 1 aa deletions versus the relative solvent accessible area (rSASA) with Pearson's correlation coefficients. **b**. Histogram of Pearson's correlation coefficients of indel or substitution stability ddG and rSASA for the Tsuboyama et al. dataset. **c.** Patterns of periodicity for substitutions, insertions and deletions across helices and **d**. strands. Number of helices/strands for each position is reported above the plot; red/blue line: mean ddG/position. **e**. Violin plots of substitution and indel stability ddG across i) termini ii) loops iii) helices and iv) strands. Significant changes are indicated by the p-values obtained through Bonferroni-adjusted Wilcox two-sided t-test.

Figure 4. Prediction of indel tolerance. a. Histograms of Pearson's correlations coefficients for observed and predicted tolerance scores of substitutions (purple), insertions (blue) and deletions (green) across the domains from Tsuboyama et al.. Prediction accuracy of CADD could only be tested on human domains. **b**. Overview of the predictive features and the indel prediction models. **c**. A regularised (lasso) multiple linear regression model for prediction of deletion and **d.** insertion tolerance. The performance was evaluated as leave-one-out cross validation and is reported as Pearson's correlation between observed and predicted scores for 178 domains from Tsuboyama et al..

Figure 5. Insertions generate gain-of-function molecular phenotypes. a. Overview of the bPCA selections; no: yeast growth defect; DHF: dihydrofolate; THF: tetrahydrofolate. **b**. Density distributions of 1-3 aa deletion (green), 1-3 aa insertion (blue) and single substitution variants (purple); grey: aPCA data. **c**. Heatmaps of 19 single aa substitutions, 20 aa insertions and 1-3 aa deletion effects on protein-protein binding. Changes in binding significant from the weighted mean of the synonymous variants are marked with "*" (Bonferroni multiple testing correction of z-stats in a two-tailed test); y-axis: identity or length of mutation; x-axis: mutated position- and the aa sequence of the domain. The x-axis is coloured by a secondary structure element; black: termini/loop; red: helix; blue: strand residues; black squares: residues directly involved in the binding interaction. **d**. Mean insertion scores/residue for bPCA (blue) and aPCA (grey); error bars: standard deviation of aPCA scores/residue. **e**. Scatter plots of 19 substitution and 20 insertion effects on protein abundance and binding; red: residues in binding surface; blue: residues with rSASA<30; green: residues with rSASA>30. **f.** Structures of domains interacting with their ligands coloured by the mean bPCA score of 20 aa insertions/residue.

Extended Data Figure 1. Experimental reproducibility and comparisons of single and multi-aa indels. a. Scatter plots showing reproducibility of abundance scores for the 9 domains. **b**. Scatter plots showing correlation of aPCA scores and inferred ddG stability scores from Tsuboyama et al., for indels across the overlapping 5 domains. Pearson's correlation reported for each domain. The table reports on differences in domain boundaries across the experiments; green: deletions; blue: insertions. **c**. Density distributions of 1 aa deletions (green) and 1 aa CCC insertions (blue) aPCA scores for all 9 domains. **d-j**. Scatter plots of 1-3 aa insertion versus deletion aPCA scores (d-f) and single versus multi-aa insertion or deletion aPCA scores (g-h, i-j respectively) across the 9 domains. Results are reported as scatter plots for all 9 domains and bar plots with Pearson's correlation coefficients for each individual domain.

Extended Data Figure 2. Substitutions and insertions across secondary structure elements. a. Scatter plots of correlations between 1 aa CCC insertions before or after the substituted residue and the mean substitution aPCA score/residue or single deletion score with their corresponding Pearson's correlation coefficients. **b**. Histograms of Pearson's correlation coefficients for deletions versus insertions before or after the deleted residue and substitutions versus insertions before or after the substituted residue. **c**. Histograms of Pearson's correlation coefficients for substitutions or deletions versus insertions before the substituted/deleted residue across secondary structure and loop residues. **d**. Protein abundance scores of 19 aa substitutions per position; blue dot: substitutions to alanine. **e**. Scatter plots of mean insertion and mean substitution scores per aa identity within loops, strands, 310helices and alpha helices. **f-g**. Scatter plots of all 19 substitutions versus 19 insertions after or before the substituted residue.

Extended Data Figure 3. Variability of indel tolerance across secondary structure features. a. Histograms of Pearson's correlation coefficients for ddG of substitutions (purple), insertions (blue) and deletions (green) versus relative solvent accessible area (rSASA) **b**. Substitution, insertion and deletion tolerance across short (1-3 aa) and long (=>4aa) n- and c-termini. Number of termini/position are indicated above the plots; x-axis: realigned termini positions. **c**. ddG scores for substitutions, insertions and deletions across different helix, strand and loop lengths. **d**. ddG scores for substitutions, insertions and deletions across helix/strands or **e.** loops classified by the secondary structure before or after the element. "start" and "end" denote the ddG of first and last structural elements of the domain.

Extended Data Figure 4. Indel variant effect prediction. a. Plots of correlations between the predicted PROVEAN and CADD substitution scores. Results are shown as a scatter plot and histogram with the per-domain Pearson's correlation; n=42. **b**. Scatter plots of the correlation between the predicted PROVEAN and CADD insertion and deletion scores and the corresponding Pearson's correlation; n=42 **c-d**. Top significant coefficients from the deletion (c) and insertion (d) predictor models 5 and 5p.

Extended Data Figure 5. Impact of multi-aa insertions and delSubs on protein interactions. a-b. Heatmaps of 1-3 aa insertion repeats and delSubs bPCA scores. Changes in binding significant from the weighted mean of the synonymous variants are marked with "*" (Bonferroni multiple testing correction of z-statistic in a two-tailed test); "subID" denotes the identity of the substitution; y-axis: type of mutation; x-axis: mutated position and identity of the substitution for delSub. The x-axis is coloured by a secondary structure element; black: termini/loop; red: helix; blue: strand residues.

Fig. 1

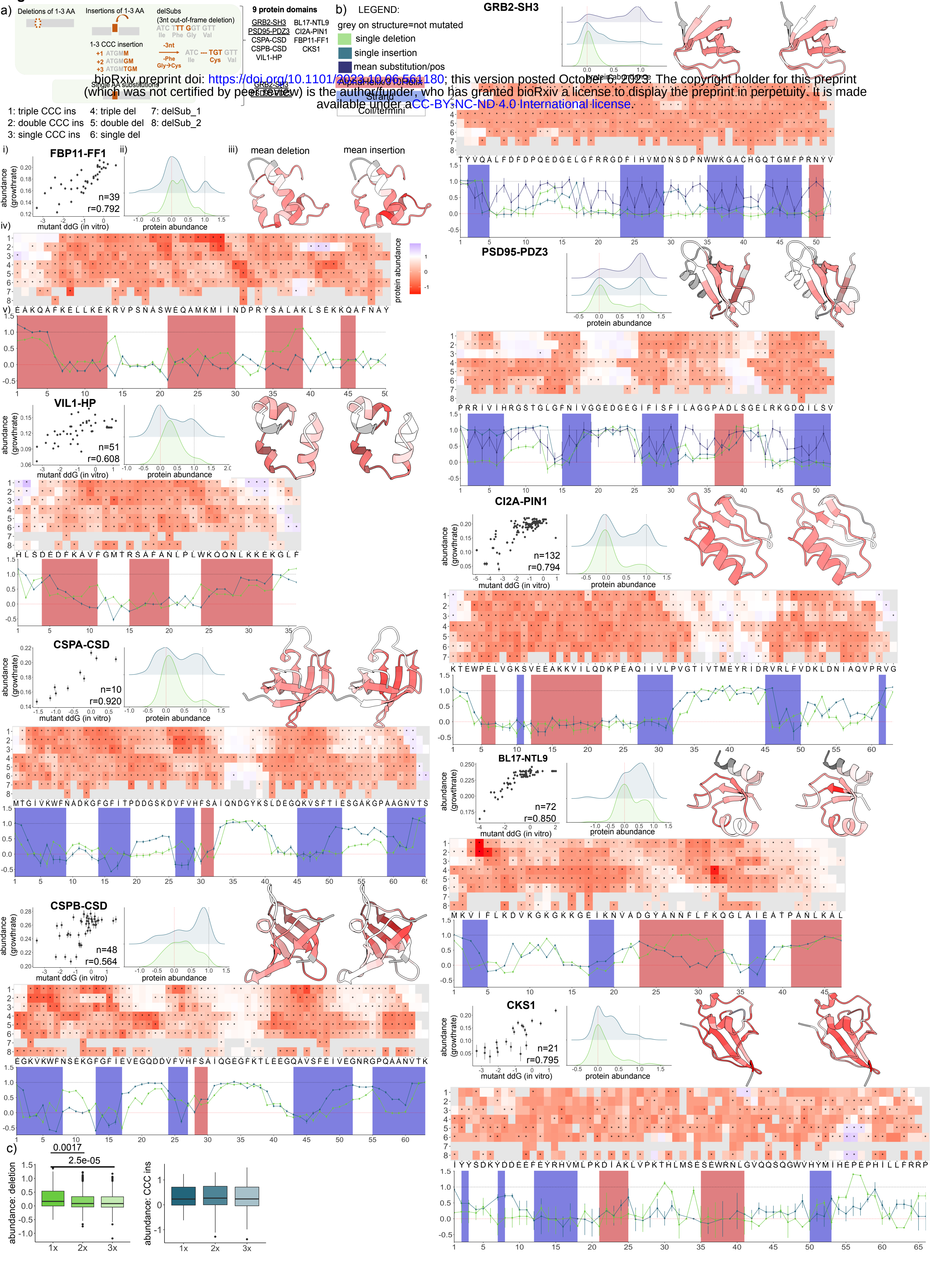

a)
Deletions of 1-3 AA
Insertions of 1-3 AA
delSubs
(3nt out-of-frame deletion)
ATC TTT GGT GTT
Ile Phe Gly Val
1-3 CCC insertion
+1 ATGMM
+2 ATGMGM
+3 ATGMTGM
-3nt
-Phe
Gly→Cys
ATC --- TGT GTT
Ile Cys Val
Single AA substitutions
9 protein domains
GRB2-SH3
PSD95-PDZ3
CSPA-CSD
CSPB-CSD
VIL1-HP
BL17-NTL9
CI2A-PIN1
FBP11-FF1
CKS1
1: triple CCC ins
2: double CCC ins
3: single CCC ins
4: triple del
5: double del
6: single del
7: delSub_1
8: delSub_2
b) LEGEND:
grey on structure=not mutated
single deletion
single insertion
mean substitution/pos
AlphaHelix/310Helix
Strand
Coil/termini
i)
ii)
iii)
iv)
v)
FBP11-FF1
abundance (growthrate)
mutant ddG (in vitro)
n=39
r=0.792
protein abundance
mean deletion
mean insertion
protein abundance
EAKQAFKELLKEKRVPSNASWEQAMKMIINDPRYSALAKLSEKKQAFNAY
VIL1-HP
n=51
r=0.608
HLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
CSPA-CSD
n=10
r=0.920
MTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAKGPAAGNVTS
CSPB-CSD
n=48
r=0.564
EGKVKWFNSEKGFGFIEVEGQDDVFVHFSAIQGEGFKTLEEGQAVSFEIVEGNRGPQAANVTK
GRB2-SH3
TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYV
PSD95-PDZ3
PRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSV
CI2A-PIN1
n=132
r=0.794
KTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAQVPRVG
BL17-NTL9
n=72
r=0.850
MKVIFLKDVKGKGKKGEIKNVADGYANNFLFKQGLAIEATPANLKAL
CKS1
n=21
r=0.795
IYYSDKYDDEEFEYRHVMLPKDIAKLVPKTHLMSESEWRNLGVQQSQGWVHYMIHEPEPHILLFRRP
c)
0.0017
2.5e-05
abundance: deletion
1x
2x
3x
abundance: CCC ins
1x
2x
3x

Fig. 2

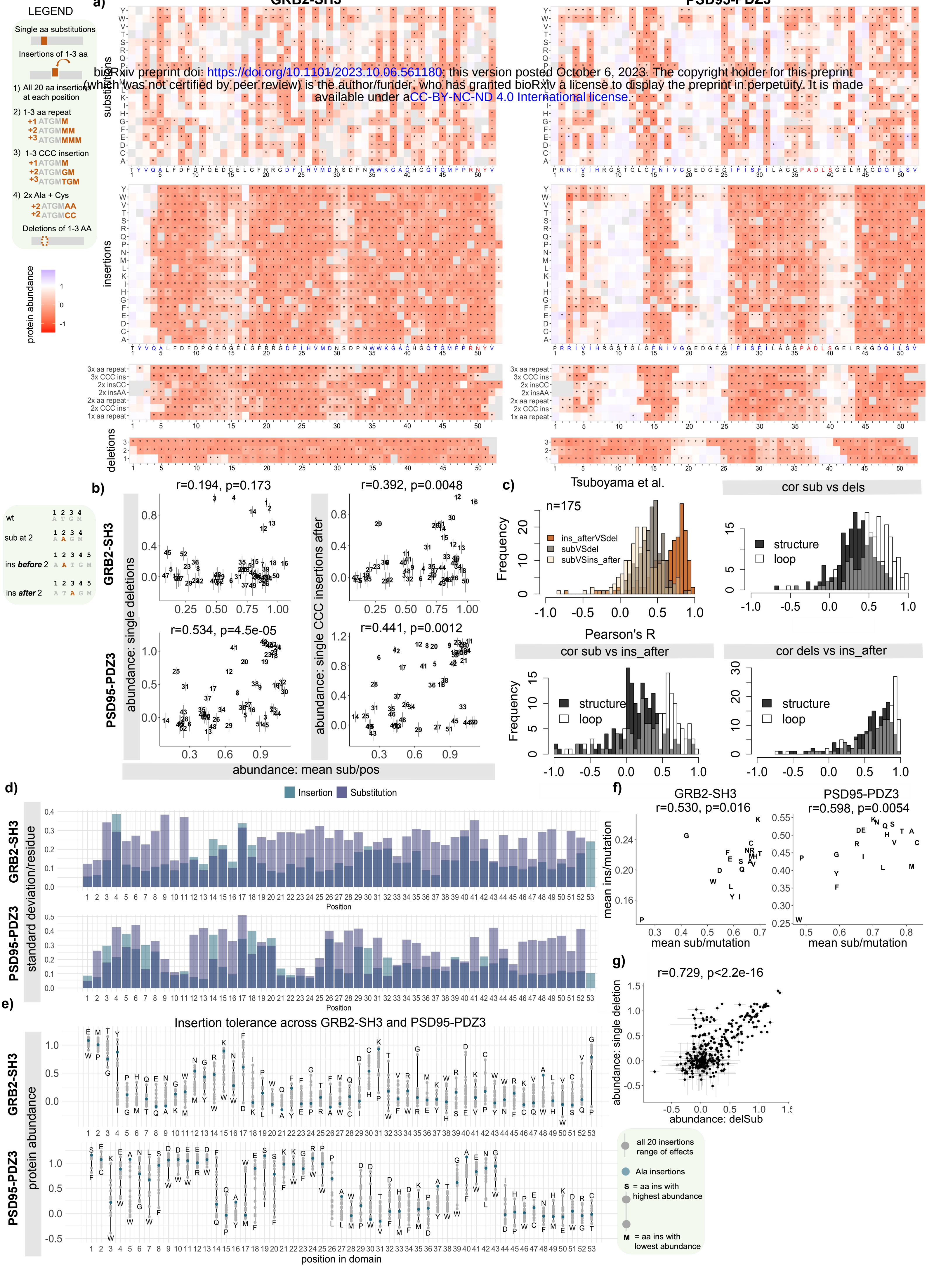

LEGEND
Single aa substitutions
Insertions of 1-3 aa
1) All 20 aa insertions at each position
2) 1-3 aa repeat
+1 ATGMM
+2 ATGMMM
+3 ATGMMMM
3) 1-3 CCC insertion
+1 ATGMM
+2 ATGMGM
+3 ATGMTGM
4) 2x Ala + Cys
+2 ATGMAA
+2 ATGMCC
Deletions of 1-3 AA
protein abundance
1
0
-1
a)
GRB2-SH3
PSD95-PDZ3
substitutions
insertions
deletions
TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYV
PRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSV
3x aa repeat
3x CCC ins
2x insCC
2x insAA
2x aa repeat
2x CCC ins
1x aa repeat
wt 1 2 3 4 A T G M
sub at 2 1 2 3 4 A A G M
ins before 2 1 2 3 4 5 A A T G M
ins after 2 1 2 3 4 5 A T A G M
b)
GRB2-SH3
PSD95-PDZ3
r=0.194, p=0.173
r=0.392, p=0.0048
r=0.534, p=4.5e-05
r=0.441, p=0.0012
abundance: single deletions
abundance: single CCC insertions after
abundance: mean sub/pos
c)
Tsuboyama et al.
n=175
ins_afterVSdel
subVSdel
subVSins_after
Frequency
Pearson's R
cor sub vs dels
cor sub vs ins_after
cor dels vs ins_after
structure
loop
d)
Insertion
Substitution
GRB2-SH3
PSD95-PDZ3
standard deviation/residue
Position
e)
Insertion tolerance across GRB2-SH3 and PSD95-PDZ3
GRB2-SH3
PSD95-PDZ3
protein abundance
position in domain
all 20 insertions range of effects
Ala insertions
S = aa ins with highest abundance
M = aa ins with lowest abundance
f)
GRB2-SH3
r=0.530, p=0.016
PSD95-PDZ3
r=0.598, p=0.0054
mean ins/mutation
mean sub/mutation
g)
r=0.729, p<2.2e-16
abundance: single deletion
abundance: delSub

Fig. 3

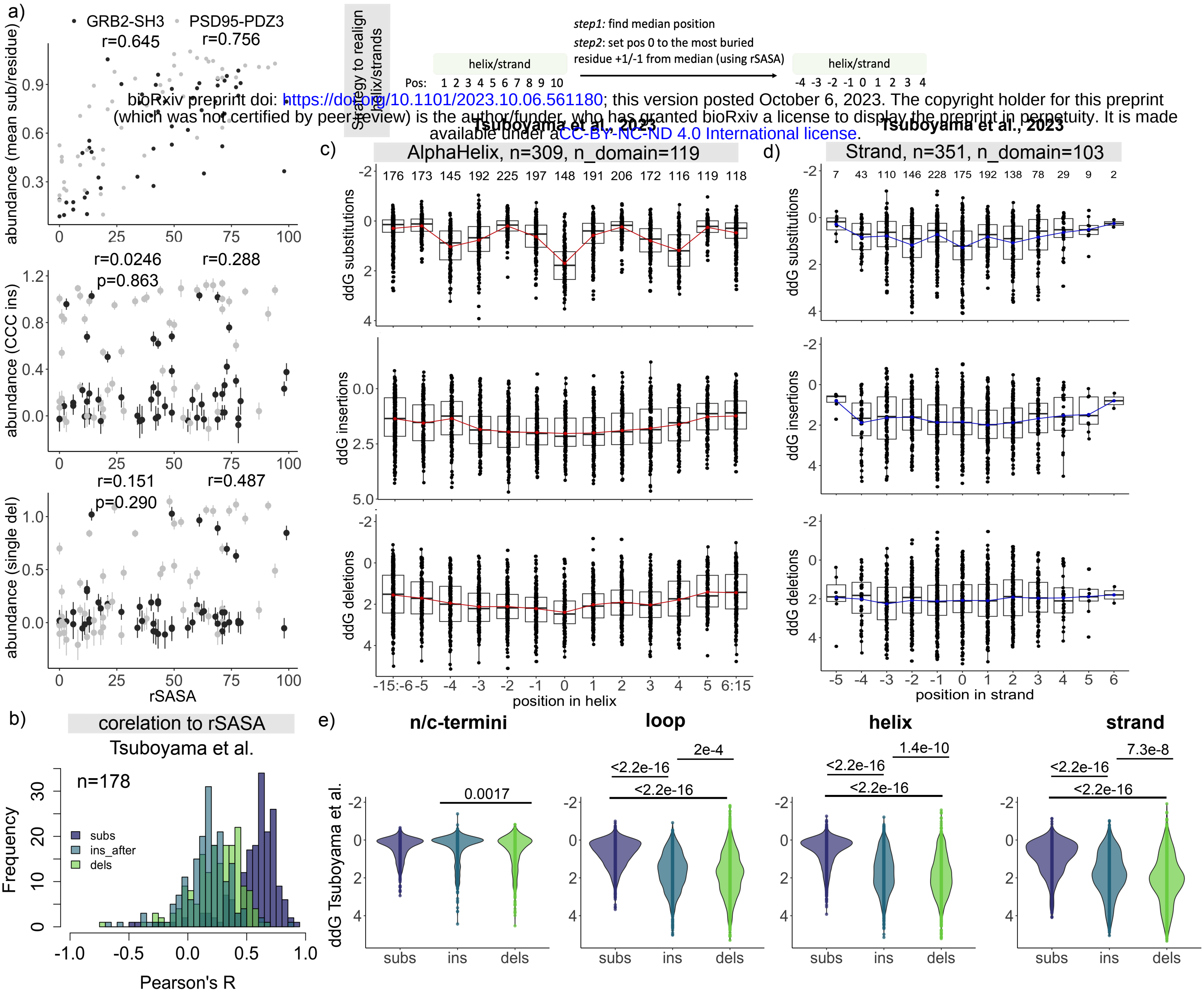

a)
GRB2-SH3
PSD95-PDZ3
r=0.645
r=0.756
abundance (mean sub/residue)
r=0.0246
p=0.863
r=0.288
abundance (CCC ins)
r=0.151
p=0.290
r=0.487
abundance (single del)
rSASA
Strategy to realign helix/strands
Pos: 1 2 3 4 5 6 7 8 9 10
helix/strand
step1: find median position
step2: set pos 0 to the most buried residue +1/-1 from median (using rSASA)
-4 -3 -2 -1 0 1 2 3 4
Tsuboyama et al., 2023
c)
AlphaHelix, n=309, n_domain=119
176 173 145 192 225 197 148 191 206 172 116 119 118
ddG substitutions
ddG insertions
ddG deletions
-15:-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6:15
position in helix
Tsuboyama et al., 2023
d)
Strand, n=351, n_domain=103
7 43 110 146 228 175 192 138 78 29 9 2
-5 -4 -3 -2 -1 0 1 2 3 4 5 6
position in strand
b)
corelation to rSASA
Tsuboyama et al.
n=178
subs
ins_after
dels
Frequency
Pearson's R
e)
n/c-termini
loop
helix
strand
0.0017
2e-4
<2.2e-16
<2.2e-16
1.4e-10
<2.2e-16
<2.2e-16
7.3e-8
<2.2e-16
<2.2e-16
ddG Tsuboyama et al.
subs
ins
dels

Fig. 4

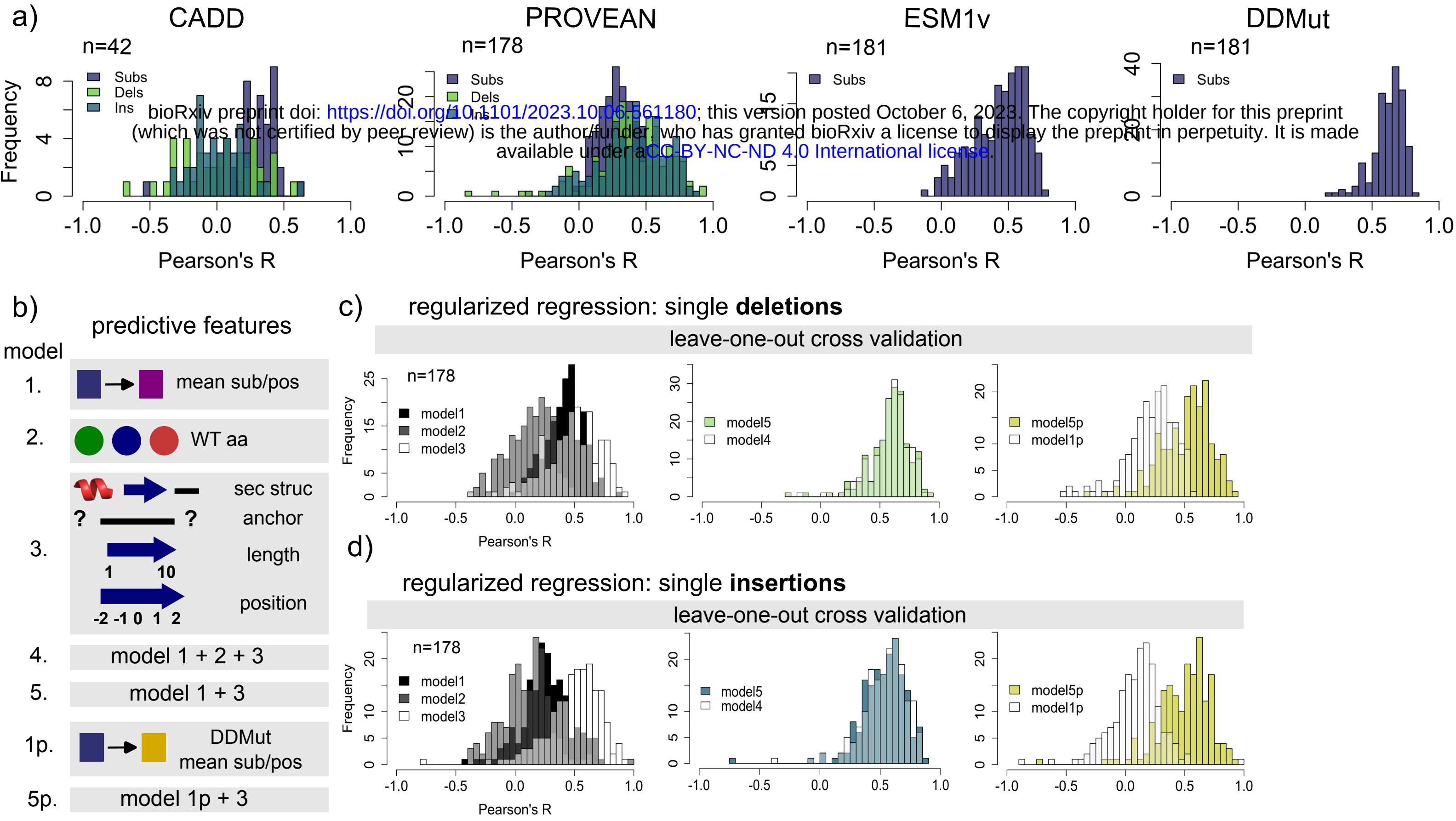
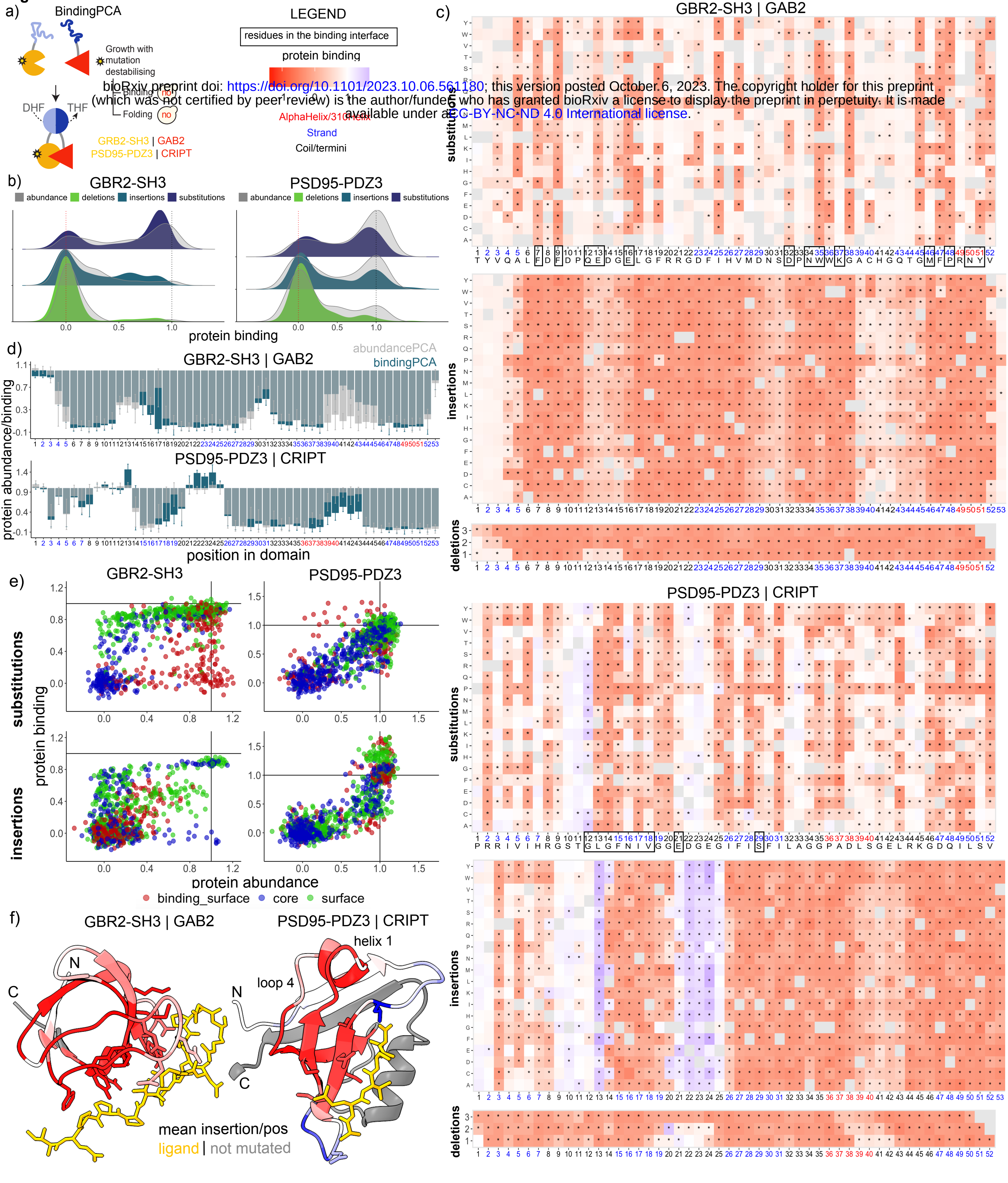
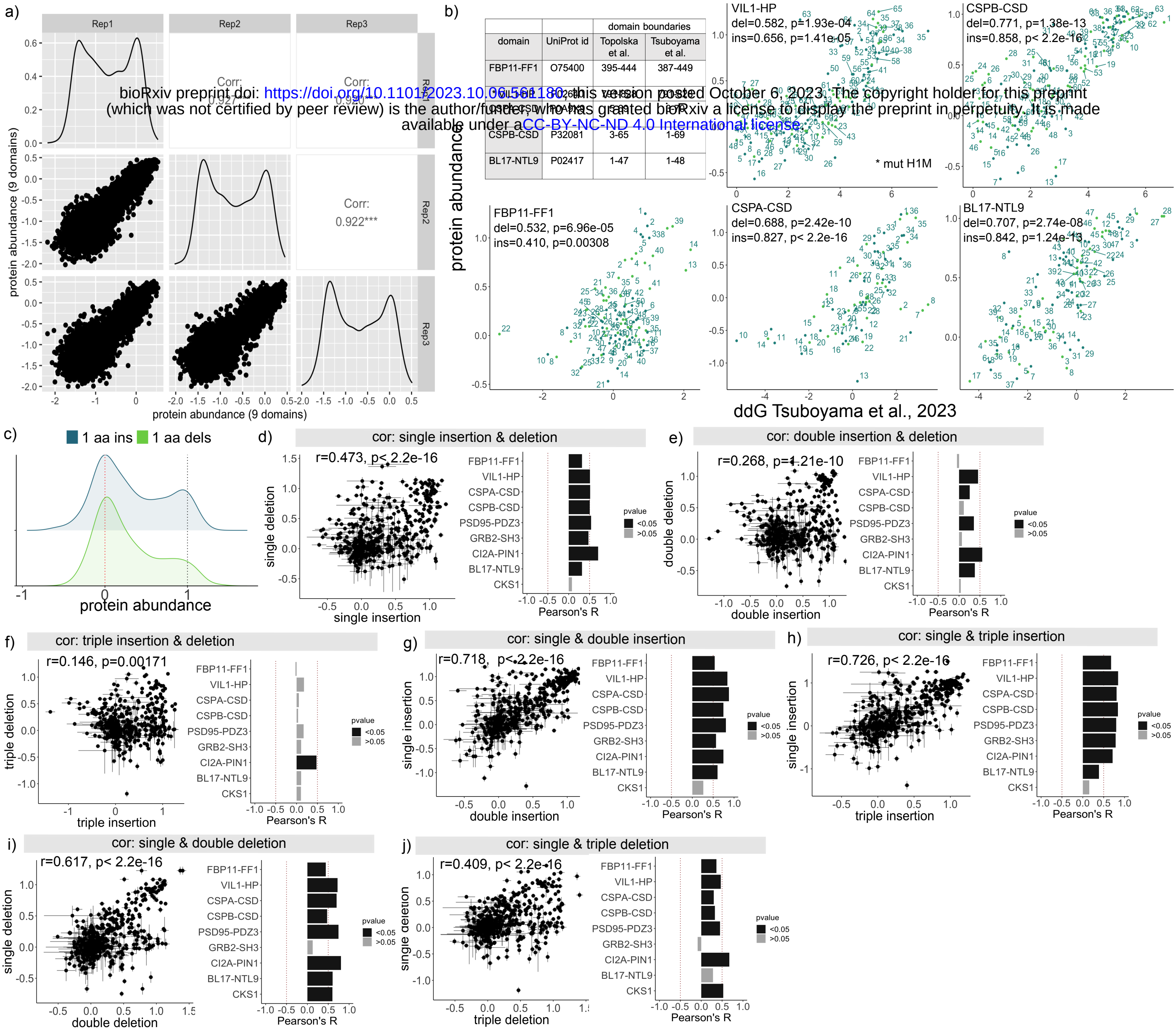

a)
CADD
n=42
PROVEAN
n=178
ESM1v
n=181
DDMut
n=181
Subs
Dels
Ins
Frequency
Pearson's R
-1.0
-0.5
0.0
0.5
1.0
b)
predictive features
model
1.
mean sub/pos
2.
WT aa
3.
sec struc
?
?
anchor
1
10
length
-2 -1 0 1 2
position
4.
model 1 + 2 + 3
5.
model 1 + 3
1p.
DDMut
mean sub/pos
5p.
model 1p + 3
c)
regularized regression: single **deletions**
leave-one-out cross validation
n=178
model1
model2
model3
model5
model4
model5p
model1p
d)
regularized regression: single **insertions**
leave-one-out cross validation
n=178
model1
model2
model3
model5
model4
model5p
model1p
Frequency
Pearson's R

Fig. 5

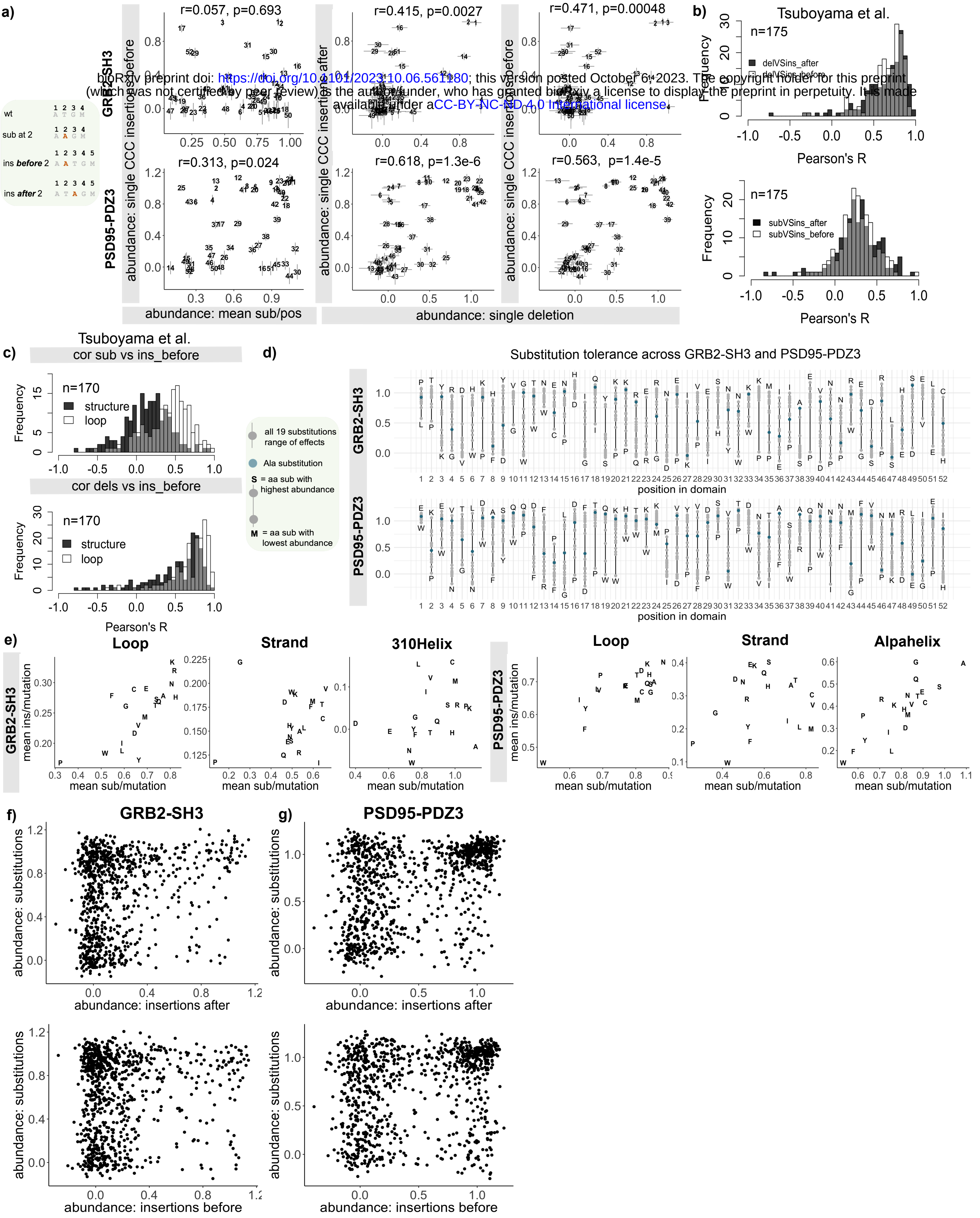

a)
BindingPCA
Growth with mutation destabilising
Binding no
Folding no
DHF
THF
GRB2-SH3 | GAB2
PSD95-PDZ3 | CRIPT
LEGEND
residues in the binding interface
protein binding
-1
0
1
AlphaHelix/310helix
Strand
Coil/termini
b)
GBR2-SH3
PSD95-PDZ3
abundance
deletions
insertions
substitutions
0.0
0.5
1.0
protein binding
c)
GBR2-SH3 | GAB2
substitutions
insertions
deletions
T Y V Q A L F D F D P Q E D G E L G F R R G D F I H V M D N S D P N W W K G A C H G Q T G M F P R N Y V
PSD95-PDZ3 | CRIPT
P R R I V I H R G S T G L G F N I V G G E D G E G I F I S F I L A G G P A D L S G E L R K G D Q I L S V
d)
GBR2-SH3 | GAB2
PSD95-PDZ3 | CRIPT
abundancePCA
bindingPCA
protein abundance/binding
position in domain
e)
GBR2-SH3
PSD95-PDZ3
substitutions
insertions
protein binding
protein abundance
binding_surface
core
surface
f)
GBR2-SH3 | GAB2
PSD95-PDZ3 | CRIPT
N
C
helix 1
loop 4
mean insertion/pos
ligand | not mutated

Extended Data Fig. 1

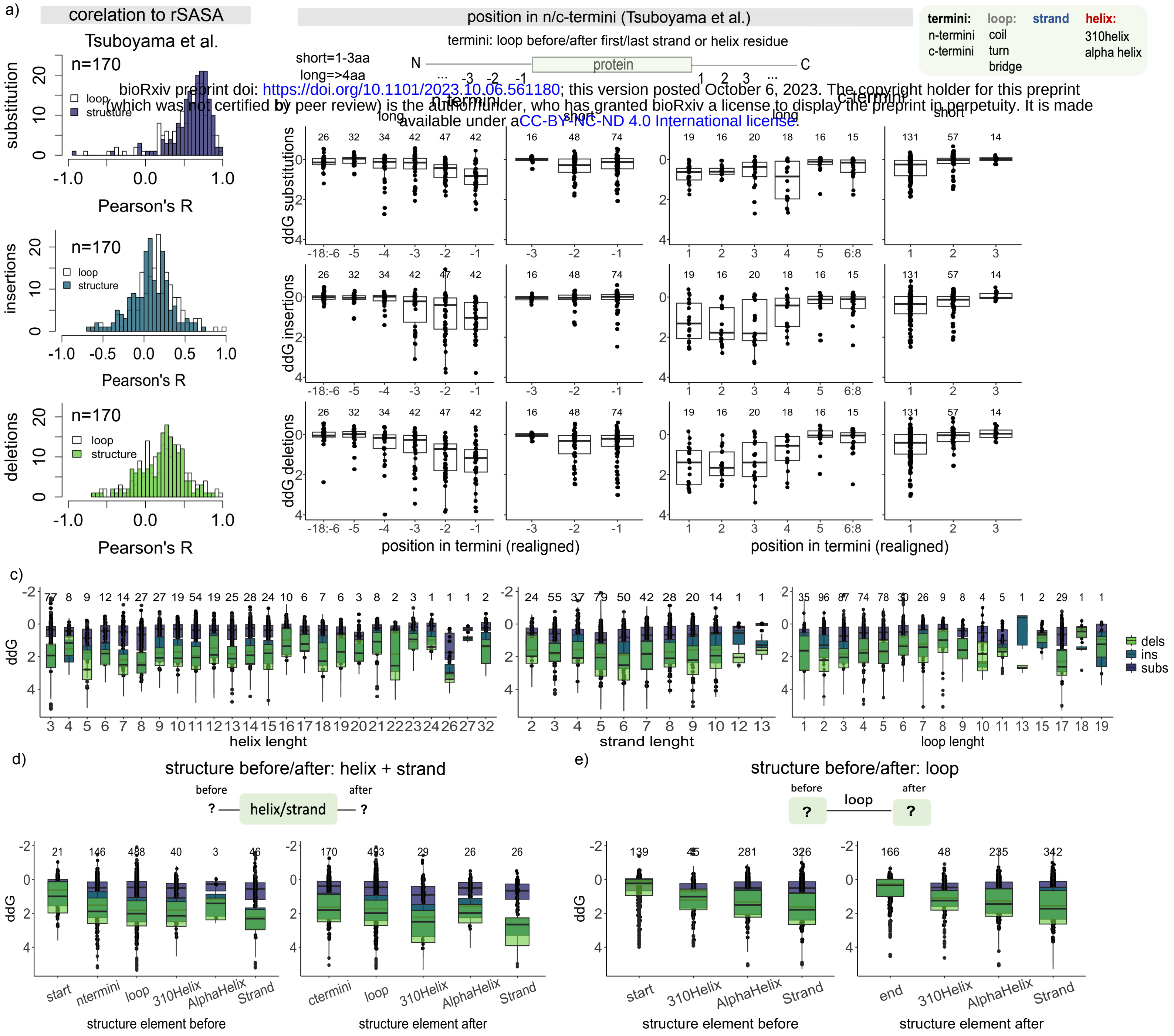


		domain boundaries	
domain	UniProt id	Topolska et al.	Tsuboyama et al.
FBP11-FF1	O75400	395-444	387-449
* VIL1-HP	P02640	791-828	791-826
CSPA-CSD	P0A9X9	5-69	3-70
CSPB-CSD	P32081	3-65	1-69
BL17-NTL9	P02417	1-47	1-48

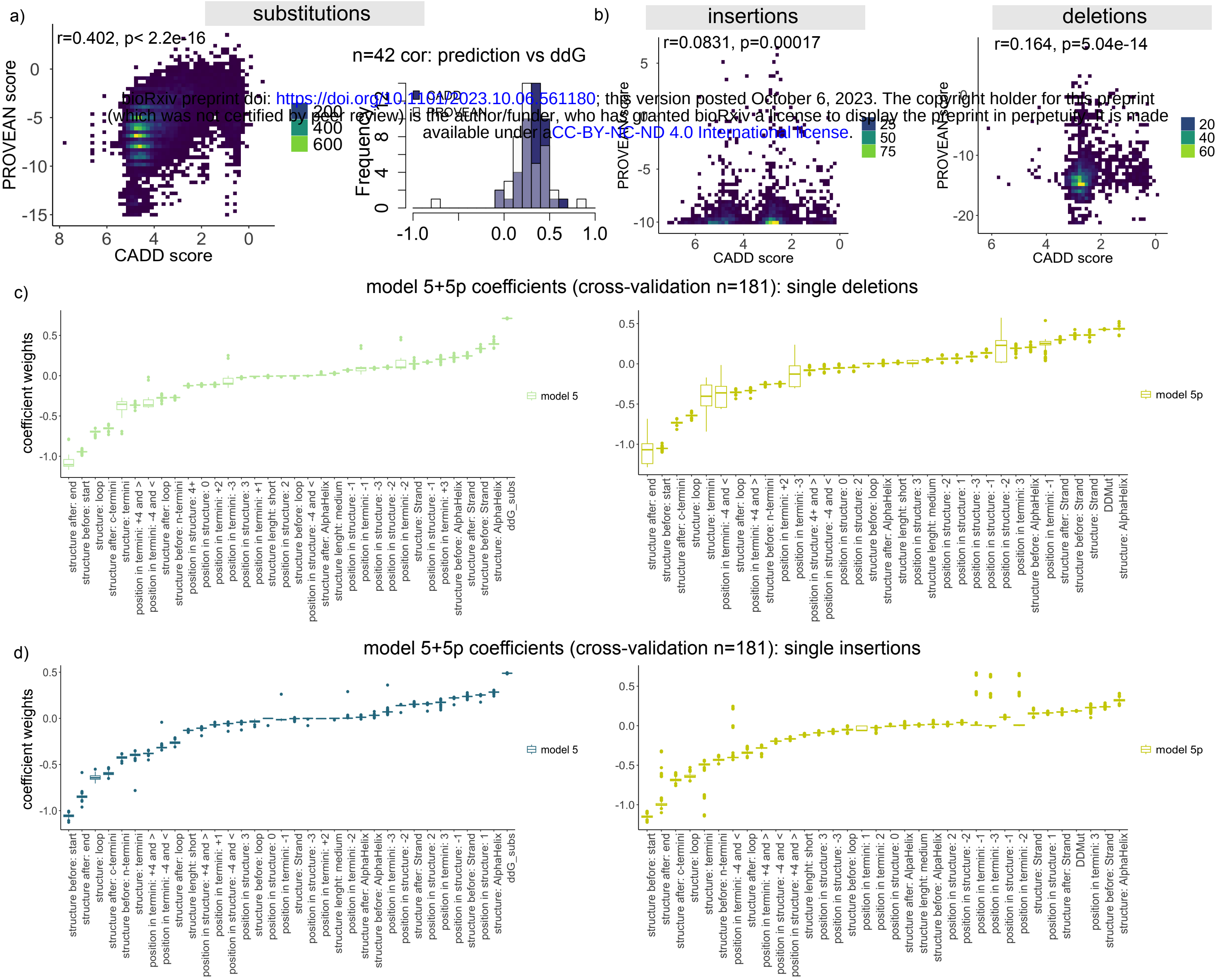
Extended Data Fig. 2



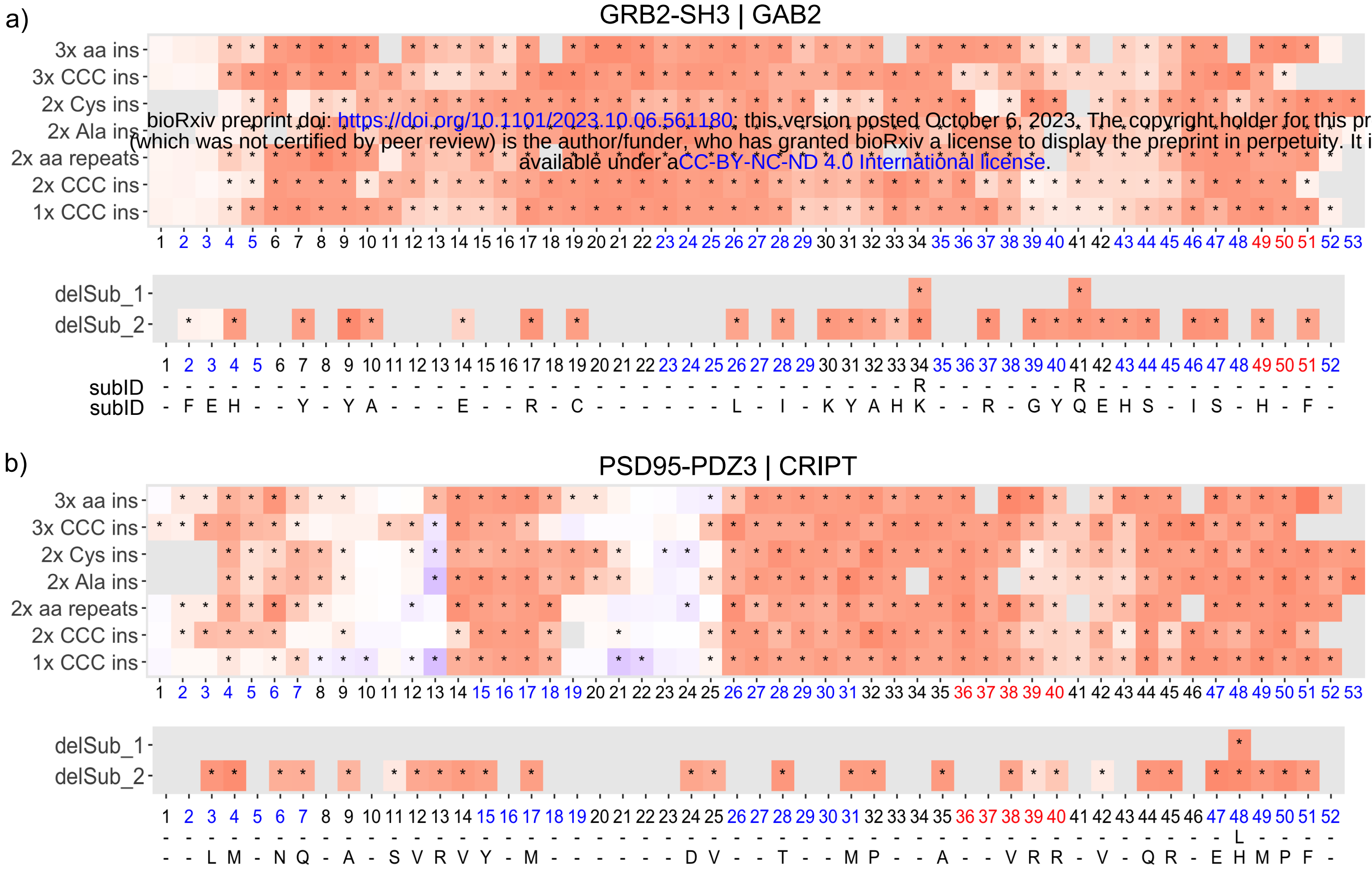
Extended Data Fig. 3



Extended Data Fig. 4



Extended Data Fig. 5


a)
GRB2-SH3 | GAB2
3x aa ins
3x CCC ins
2x Cys ins
2x Ala ins
2x aa repeats
2x CCC ins
1x CCC ins
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
delSub_1
delSub_2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
subID - R - - - - - - R - - - - - - - - - - -
subID - F E H - - Y - Y A - - - - E - - R - C - - - - - - - - L - I - K Y A H K - - R - G Y Q E H S - I S - H - F -
b)
PSD95-PDZ3 | CRIPT
3x aa ins
3x CCC ins
2x Cys ins
2x Ala ins
2x aa repeats
2x CCC ins
1x CCC ins
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
delSub_1
delSub_2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
- L - - - -
- - L M - N Q - A - S V R V Y - M - - - - - - - D V - - T - - M P - - A - - - V R R - V - Q R - E H M P F -