Data used in comparisons among SAAPs, wt-DARs, and rg-DARs

SAAPs from UniProt are listed in *saap_maf.list*. For each SAAP, the file provides the UniProt ID of the protein associated with the SAAP, position of the SAAP in the protein sequence, reference allele and minor allele for the SAAP, and the minor allele frequency (MAF). The file *saap_maf001.ali* contains alignments of protein sequences and their structures associated with the common SAAPs (MAF≥0.01). Positions of the SAAPs in structures are identified from the alignments. The headline of an alignment in the file records the protein UniProt ID, structure ID, SAAP position in the structure, reference allele, and the minor allele. File *saap_maf001.ali* was used for Figure 1.

All DARs are listed in *dar.list*. For wt-DARs, the file *wt-DAR.ali* includes alignments of human protein sequences, structures, and orthologs from species where the wt-DARs appear as wild-types. The alignment headline includes protein UniProt ID, structure ID, and wt-DAR position in the structure. It also contains the information of substitutions between aligned residues in the neighborhood of the wt-DAR site, including the substitution number, types, and positions in the structure. For rg-DARs, the file *rg-DAR.ali* contains alignments between protein sequences and their structures associated with the rg-DARs. In the file, the headline of an alignment has the protein UniProt ID, structure ID, rg-DAR position in the structure, and the wild-type residue. The data in *rg-DAR.ali* and *wt-DAR.ali* were used for Figure 1.

Data used in comparisons among wt-DARs, wt-DARs with potential compensatory residues, and non-compensatory residues

The file *compensated_wt-DAR.list* is a list of wt-DARs generated from *wt-DAR.ali*. Each wt-DAR in this subset satisfies the criteria that in the structures-ortholog sequence alignment associated with it, none of the neighboring residues of the wt-DAR site corresponds to a gap site in the ortholog and at least one neighboring residue differs between the structure and the ortholog. The file *compensated_wt-DAR.ali* contains alignment and substitution information for this subset of wt-DARs. The headline of an alignment in the file records the protein UniProt ID, structure ID, wt-DAR position in the structure, the number of substitutions detected in the neighborhood of the wt-DAR site, the substitution types and locations. The file *noncompensated_wt-DAR.ali* contains alignments of human protein sequences, structures, and orthologs that have the same reside as human wild-type residues at wt-DAR sites. The headline of an alignment in the file records the protein UniProt ID, structure ID, wt-DAR position in the structure, the number of substitutions detected in the neighborhood of the wt-DAR site, the substitution types and locations. Data in *compensated_wt-DAR.ali* were used for Figure 3 and Figure S3.

From *compensated_wt-DAR.ali* and *noncompensated_wt-DAR.ali*, wt-DARs with the same numbers of potential compensatory residues and non-compensatory residues are identified. Their alignment and substitution information are deposited in a zipped file *comp_noncomp_equal.tar.gz*. Data in the files were used for Figure 4.

The file *compensated_wt-DAR.ali.seqid60* contains a non-redundant subset collected from *compensated_wt-DAR.ali* by CD-Hit [1] to ensure that the structures in the subset have sequence identity $\leq$ 60%. The file was used for Figure S4.

Data used in detecting specificity of potential compensatory residues

This data set was generated from *compensated_wt-DAR.ali* as described in the Materials and Methods section. In the file *DARLvsWTRL.xlsx*, each wt-DAR has protein ID, position in structure, wild-type residue, DARL(s), and WTRL(s). For a mutation from wild-type to a DARL, the $\Delta\Delta$Gs were calculated with the presence and absence of the potential compensatory residues, respectively. $\Delta\Delta$Gs were also calculated for mutations from wild-types to WTRLs. These data were used for Figure 5.

Protein structures and residue distances

All protein structures used in the paper are deposited in pdb.tar.gz, including native structures and models. For each structure, minimum distance of non-hydrogen atoms between residues were calculated by *mindist.pl* in the MMTSB toolset [2].

1. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17: 282-283.
2. Feig M, Karanicolas J, Brooks CL, 3rd (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of molecular graphics & modelling 22: 377-395.