Supporting Information

S1 Datasets of non-synonymous SNVs & their structural coverage

In order to evaluate the impact of various types of non-synonymous variants on localized frustration of protein residues in different biological context, we collected and analyzed data from a variety of sources. These sources were chosen in order to obtain both benign and disease-associated SNVs, with the disease-associated SNVs having been further sub-classified to investigate their associated mode of action in greater detail. An overview of this data collection scheme is provided in ***S1A***. ***S1B*** gives summary statistics on all non-synonymous SNVs in these datasets, and ***S1C*** provides the corresponding data on the subset of these SNVs, which were mapped to high-resolution protein structures from the PDB. Further details on the statistics obtained as part of this data collection framework is provided below.

We collected and annotated 6.46 million non-synonymous SNVs using VAT. About 5.1 million of these SNVs were benign mutations that were obtained from the ExAC Project, and an additional roughly 0.6 million SNVs were taken from phase 3 of the 1000 Genomes Project, which constitutes 79% and 9% of our total set of annotated SNVs, respectively (***S1B***). The remaining SNVs were a set of disease-associated mutations, and these comprised ~76,000 HGMD SNVs and 0.65 million publicly available pan-cancer somatic SNVs. HGMD and the pan-cancer dataset constituted 2% and 10% of the total collected non-synonymous SNVs, respectively (Figure ***S1B***).

However, the contribution of SNVs from different resources changed significantly while considering only those annotated SNVs, which mapped to high-resolution protein structures. Approximately 96,000 SNVs from ExAC were mapped to protein structures in the PDB constituting 51% of our totals set of structurally mapped SNVs (Figure ***S1C***). Similarly, 1KG SNVs constituted 7% (13588) of the total structurally mapped SNV dataset. In contrast, the percentage of the disease-associated SNVs that were mapped to protein structures was 18% (33,261 SNVs) and 24% (44,094 SNVs) for the HGMD and pan-cancer resource, respectively (Figure ***S1C***). The majority of SNVs from the pan-cancer dataset that were mapped to protein structures impacted cancer-associated genes (CAG), constituting 14% (25,409) of the all SNVs mapped to protein structure, whereas SNVs impacting non-cancer associated genes constituted only

8% (15,044) (Figure ***S1C***). In contrast, 4,041 SNVs affecting driver genes mapped to protein structures; these SNVs constitute 2% of the total structurally mapped non-synonymous SNVs (Figure ***S1C***).

S2 Frustration differences for semi-balanced structure datasets

Deleted: **Variant statistics on**

After identifying unique protein sets, those proteins which fall within the *intersection* of the different datasets were used to evaluate ΔF distributions. For instance, the intersection between unique 1000 Genomes and HGMD proteins constitutes a non-redundant set of protein structures in which at least one residue intersects with a 1000 Genomes SNV and at least one residue intersects with an HGMD SNV, thereby providing a semi-balanced set of SNVs, and thus the ability to draw more direct comparisons with respect to ΔF distributions. Using this approach, we find that the results overall trend for semi-balanced variant datasets are consistent with ΔF distributions detailed above (Supp. Fig S5). However, they lack statistical significance, potentially due to lower amount of SNVs included in the semi-balanced dataset.

Deleted: Considerable annotation disparities exist between HGMD variants and variants taken from 1000 Genomes and ExAC, raising the possibility of bias between the evaluated structure datasets. The sets of proteins evaluated in the context of HGMD variants may thus be considerably different from those of 1000 Genomes/ExAC SNVs, thereby making direct comparisons difficult. Finally, some protein structures may be over-represented in any given dataset. To control for these effects, we first identify a non-redundant set of unique proteins within each dataset. Specifically, the non-redundant set is constructed by ensuring that no protein within the set shares more than 90% sequence identity with any other protein in the set. We find that there are 618, 907, and 303 distinct proteins within the set of high-resolution structures impacted by 1000 Genomes, ExAC, and HGMD SNVs, respectively. Distributions delineating the number of SNVs within these unique (i.e., non-redundant) protein sets are given in Supp. Fig. S2-S4.

Deleted: these

Formatted: Indent: First line: 0"

S3 Threshold to identify potentially deleterious SNVs

As discussed in result section of the main text, disease-associated SNVs from HGMD generally induce more negative ΔF values relative to benign SNVs. Given a newly discovered SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious? We address this issue empirically by optimizing a function *f(x)* defined by two distributions(1):

Formatted: Indent: Left: 0", Hanging: 0.44", No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Deleted: We address this issue empirically by optimizing a function *f(x)* defined by two distributions:

$$f(x) = h(x) + e(x)$$

Let ΔF_{HGMD} denote the distribution of ΔF scores induced by HGMD SNVs. *h(x)* is defined to be the difference between the fraction of ΔF_{HGMD} scores less than *x* (*fract*[$\Delta F_{HGMD} < x$]) and the fraction of ΔF_{HGMD} scores greater than *x* (*fract*[$\Delta F_{HGMD} > x$]):

$$h(x) \;=\; fract[\Delta F_{HGMD} < x]) \;-\; fract[\Delta F_{HGMD} > x])$$

With ΔF_{ExAC} similarly defined for the distribution of ΔF values associated with ExAC SNVs:

$$e(x) \;=\; fract[\Delta F_{ExAC} > x]) \;-\; fract[\Delta F_{ExAC} < x])$$

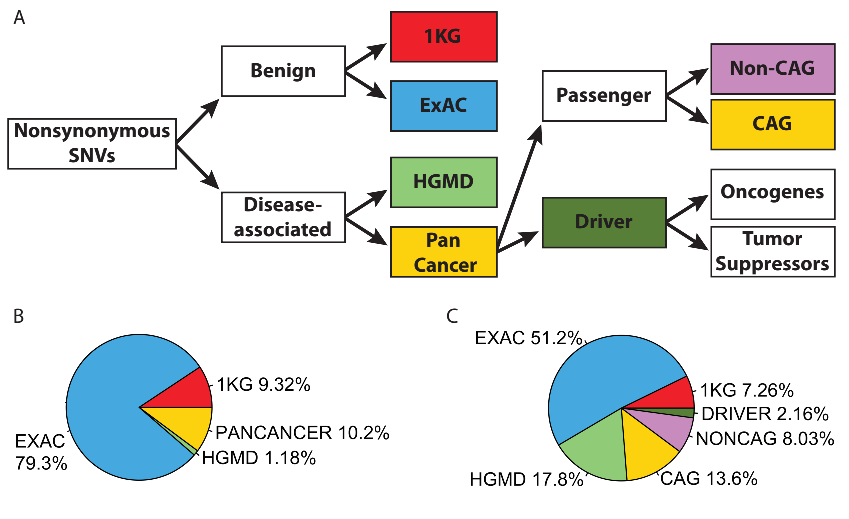
Note that, in building the distribution of ΔF_{HGMD} values, a random sample of HGMD SNVs was chosen in order to match the number of SNVs in the ΔF_{ExAC} distribution. The *x* that maximizes the function *f(x)* is taken as the ΔF threshold for predicting whether a newly discovered SNV is deleterious or benign. Using this approach, we find that this ideal threshold takes a value of $\Delta F = -1.221$.

Formatted: Indent: First line: 0"

Deleted: **S4 Usefulness of localized frustration approach** ... [1]

Figure S1

Figure S1: Overview of SNV categories and their relative proportions within the data pool analyzed. *A)* Flowchart representing the different categories and origins of the variants analyzed in this study. A given non-synonymous SNV can be classified as benign or disease-associated on the basis of its provenance (i.e., whether it is taken from 1000 Genomes, ExAC, HGMD or Pan-cancer variant datasets). *B)* Relative proportions of SNVs from various datasets *prior to* mapping SNVs to high-resolution PDB structures. *C)* Relative proportions of SNVs from various datasets *after* mapping SNVs to high-resolution PDB structures.



Formatted: Font:Arial Black, 14 pt, Bold, Underline, Font color: Auto

Figure S2

Figure S2: Frequency chart of the number of 1KG SNVs against the #of unique proteins

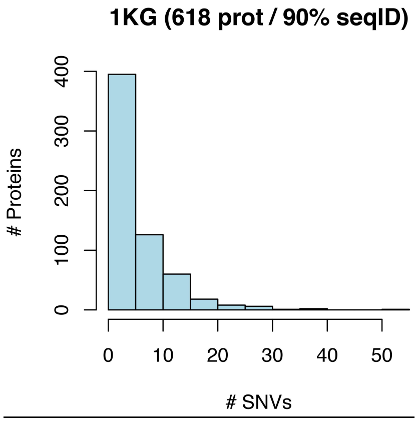
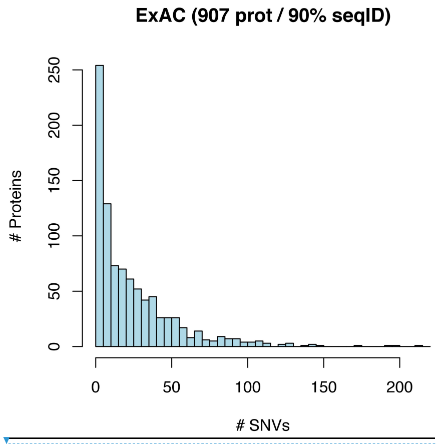


Figure S3

Figure S3: Frequency chart of the number of ExAC SNVs against the # of unique proteins.



Deleted: ... [2]

Formatted: Normal (Web)

Formatted: Font:Times New Roman, 12 pt, Not Bold, No underline, Font color: Black

Deleted:

Figure S4

Figure S4: Frequency chart of the number of HGMD SNVs against the # of unique proteins.

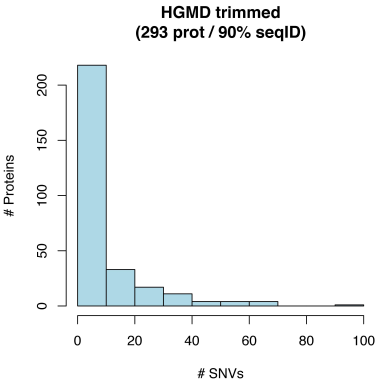


Figure S5

Figure S4: Comparison of frustration changes for circular SNVs impacting PDB structures

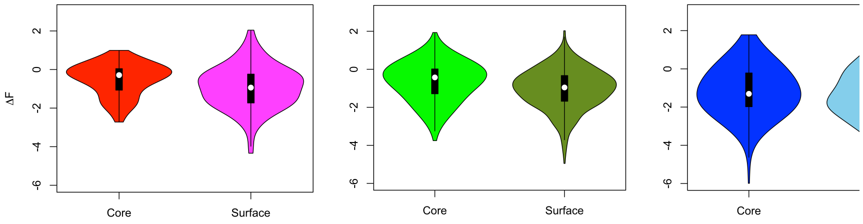


Figure S6

Figure S6: Empirical distribution to identify deleterious SNVs

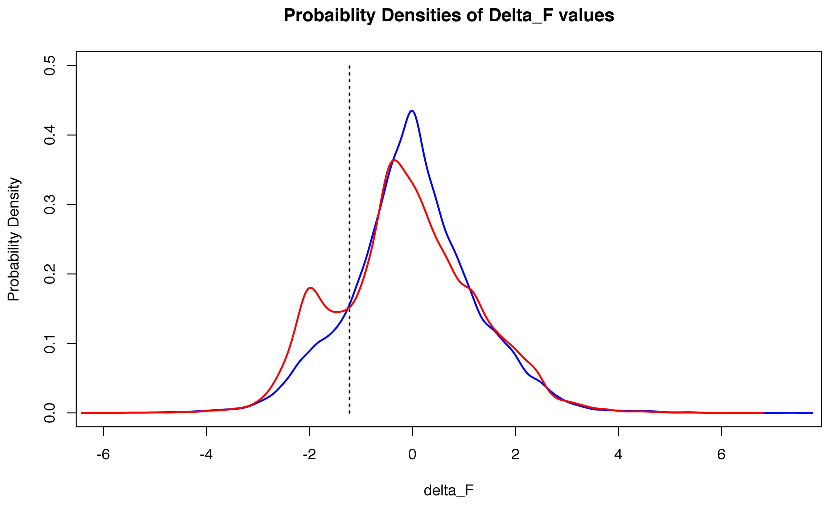
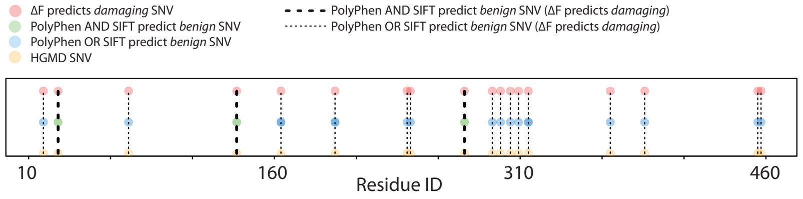


Figure S7

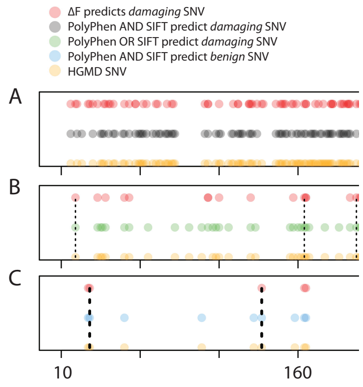
Figure S7: Example of False negatively annotated HGMD variants mapping onto protein structure by Polyphen2 & SIFT



Formatted: Font:Arial Black, 16 pt, Bold, Underline

Reference

1. Hourai Y, Akutsu T, Akiyama Y (2004) Optimizing substitution matrices by separating score distributions. *Bioinformatics* 20(6):863–73.

Deleted: 

Formatted: Font:Arial Black, 16 pt, Bold, Underline

Page 3: [1] Deleted Microsoft Office User 7/30/16 10:14 PM

S4 Usefulness of localized frustration approach

We selected a smaller set variant mapped onto PDB structure, which has at least one HGMD and at least one ExAC non-synonymous SNVs. Subsequently, we identified instances where HGMD variants were predicted to be benign by polyphen2 or SIFT (False negative cases) but delta frustration metric indicates significant increase in frustration level upon mutation. We observed that 10% of the variants in this smaller set of variants were annotated as benign by polyphen2. Similarly, SIFT incorrectly predicted 13.7% of these HGMD variants to be not damaging. Furthermore, we analyzed the delta frustration values for variants in this dataset. Applying the delta frustration threshold described earlier, we observed that 38% of the miss-annotated variants had significantly large frustration change indicating their potential deleteriousness. Furthermore, We also identified that ~46% of SIFT annotated false negative variants had large delta frustration values associated with them. We also highlight an example by plotting linear diagram for such case in the supplementary Figure S6.

Page 4: [2] Deleted Microsoft Office User 7/30/16 10:14 PM