RESPONSE LETTER

Reviewer #1

-- Ref 1.0 – annotation source, false negatives --

Reviewer Comment	It seems all SNVs curated from various resources are non-synonymous as shown in Figure S1A, but this is not clearly mentioned in the Methods part of the main text. Was the basic annotation (non-synonymous or synonymous) of all SNVs from different sources done by the VAT? As we all know, not all variants from 1KG and ExAC are "benign", did you apply any filter to minimize the potential false negatives (e.g., in silico prediction tools)?
Author Response	We would first like to thank the reviewer for taking to time to carefully read through our study, and we also thank the reviewer for valuable suggestions on how we may improve this work. In the revised version of the manuscript, we now specify that we exclusively look at non-synonymous SNVs. The reviewer has correctly pointed out that the annotations of coding SNVs were obtained using VAT. Furthermore, we agree with the reviewer that not all variants from 1KG and ExAC are necessarily benign. In an effort to deal with this, we have removed any known disease-associated variants (HGMD and TCGA) that were initially present in the 1KG and ExAC datasets.
Excerpt From Revised Manuscript	"In order to avoid redundancy and false positive call sets, we only consider HGMD SNVs annotated as pathological variants (labeled as "DM") in the HGMD dataset. Furthermore, we removed HGMD variants present in the 1000 Genomes and ExAC datasets. Similarly, we also removed known TCGA variants present in the original ExAC SNV datasets."

-- Ref 1.1 –core/surface residue description--

Reviewer Comment	Please briefly define and compare the "core" and "surface" residuals in the main text as they are critical to understand the differential impact evaluated in this study.
Author Response	We thank the reviewer for this suggestion, and we have now accordingly provided this information in the main text instead of the supplement.
Excerpt From Revised Manuscript	"Moreover, we sub-classify each of these three categories into core and surface residues based on their RSASA value. We calculated the RSASA value for each residue using NACCESS (1). Residues were defined as core when the RSASA value was lower than or equal to 25 % and surface residues had RSASA value greater than 25%."

-- Ref 1.2 – SNV frequency summary --

Reviewer Comment	Please summarize the number of SNVs used in each of your comparison analysis as Table 1 (e.g., benign/disease-associated, common/rare, conserved/variable,

	driver/passenger).
Author Response	These statistics are indeed valuable to know, and they are now provided in Table 1, which may be found within the main text.
Excerpt From Revised Manuscript	**Table 1.** Summary statistics on the number of SNVs used in comparative analyses. Shown are variant counts for non-disease *(top)*, HGMD *(bottom-left)*, and pan-cancer SNVs *(bottom-right)*.

Conservation measure	1000 Genomes core	1000 Genomes surface	ExAC core	ExAC surface
DAF rare (common)	2267 (85)	1570 (106)	17972 (102)	11550 (83)
GERP conserved (variable)	1552 (287)	1132 (212)	12165 (2174)	7637 (1406)

Conservation measure	HGMD core	HGMD surface
GERP conserved (variable)	5158 (961)	1113 (221)

SNV type	PANCAN core	PANCAN surface
non-CAG	2153	1848
CAG	4140	2767
driver	877	486

-- Ref 1.4 –variants with unknown significance --

Reviewer Comment	The results are interesting. However, I was looking forward to seeing how the workflow was applied to variants of unknown significance to help classify/predict their impact, e.g., using a certain value of ΔF as a threshold. This would be extremely valuable and useful for other investigators.
Author Response	We agree that greater value may be derived from ΔF if a specific threshold may be used when making predictions on newly discovered SNVs. In order to rigorously define a ΔF that may optimally be used to distinguish between deleterious and benign SNVs, we have taken the empirical approach of jointly analyzing the distributions of ΔF scores for HGMD (disease-associated) and SNVs from ExAC (presumably benign). The details and results of this analysis are now included within the Supplementary Materials (Supplementary text S3)
Excerpt From Revised Manuscript	*Excerpt in the Main Text* The deleteriousness of an SNV is a continuous variable, and indeed, this is reflected in the continuous nature of ΔF values. However, there is still considerable value in applying a binary classification scheme to newly discovered SNVs, which may be predicted to be benign or deleterious. In order to perform such binary classification, we applied a simplified decision boundary scheme, wherein we analyzed ΔF distributions for HGMD variants (disease-associated) and SNVs from ExAC (seemingly benign). The threshold was set with the objectives of a) minimizing the fraction of HGMD SNVs with ΔF values above the threshold, and b) minimizing the fraction of ExAC SNVs with ΔF values below the threshold. Using this approach, we observed that variants with ΔF score <= -1.221 can be considered deleterious. Details of this scheme are provided as part of the supporting information. *Excerpt in the Supplement:* As discussed in the results 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 (Supplementary figure S5) $f(x) = h(x) + e(x)$

	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_{ExAC} values, a random sample of ExAC SNVs was chosen in order to match the number of SNVs in the ΔF_{HGMD} 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 ΔF = -1.221.

-- Ref 1.5 – Clarification regarding p-value --

Reviewer Comment	In the last paragraph of Results: Differential effects of benign and disease-associated SNVs on ΔF profiles, you stated that "In addition, disease-associated SNVs (from HGMD) result in similar frustration changes between core and surface residues (p-value < 2e-16 from two-sample Wilcoxon test) (Figure 2C)." The frustration changes are similar between core and surface residues, but the p-value looks so significant (2e-16). Please confirm.
Author Response	We agree with reviewer that ΔF values are similar for the above-mentioned comparisons. Unfortunately, the p-value statement was misplaced in the original text, which was intended for the next statement (describing the comparison between HGMD core and 1KG/EXAC core residues). This has now been corrected.
Excerpt From Revised Manuscript	"However, SNVs from HGMD that impact minimally frustrated core residues induce stronger perturbations than benign SNVs influencing minimally frustrated core residues (*p-value < 2e-16 from two-sample Wilcoxon test*)"

-- Ref 1.6 – Clarification regarding p-value --

Reviewer Comment	In the last paragraph of Results: Differential effects of SNVs on oncogenes and tumor-suppressor genes, you stated that "We observed that SNVs affecting TSGs induce stronger perturbations in minimally frustrated core residues relative to surface residues (p-value = 8.15e-2 from two-sample Wilcoxon test) (Figure 6A)." It seems the difference was not significant (p = 0.08), so were you able to make this conclusion?
Author Response	Reviewer correctly points out that two-sided Wilcoxon test p-value is higher than 0.05. However, due to smaller sample size for the TSG and oncogene datasets, we also performed KS test for this comparison. KS test is considered to be very sensitive test, as it examines shape, range and median value of distribution. The two-sided KS test (p-value = 0.004765) indicate statistically significant difference between TSG and Oncogene frustration change distribution. As our conclusion was based on KS test, we update the corresponding text to corroborate this point.
Excerpt From	"We observed that SNVs affecting TSGs induce stronger perturbations in minimally frustrated

Deleted: Considering that the Wilcoxon test is known to be an underpowered test, we feel that the KS test is justified here.

Revised Manuscript	core residues relative to surface residues (*p-value = 0.004765 from two-sample KS test*) (Figure 6A)."

Deleted: *4.259e-2*

Deleted: *one-sided Wilcoxon*

-- Ref 1.7 – Fixing typographical & grammatical errors --

Reviewer Comment	There are minor misspellings or formatting errors: (a) in Methods: SNV Datasets, "Human Genome Mutational Database" should be "Human Gene Mutation Database"; (b) in Methods: Workflow to calculate frustration paragraph 2, please use the full name of "PDB" when it was first present; (c) in Discussion paragraph 1 first sentence and paragraph 4 first sentence, "…have/has lead to…" should be "…have/has led to…"; (d) in Discussion paragraph 3 third sentence, "…have grater impact…" should be "…have greater impact…"; (e) in Discussion last paragraph the next to the last sentence, "…the affects of…" should be "…the effects of…".
Author Response	We thank the reviewer for pointing out these formatting errors. They have now been corrected.
Excerpt From Revised Manuscript	"Disease-associated dataset included SNVs from the Human Gene Mutational Database (HGMD) (5) and pan-cancer dataset (45) comprising of publicly available somatic SNVs from The Cancer Genome Atlas (TCGA)" "We then integrated VAT annotation with the biomart (53) derived human gene and transcript IDs to map the SNV on to specific protein databank (PDB) structures" "In the last decade, tremendous improvements in sequencing and structural biology techniques have led to growth in genomic variation and three-dimensional structural data for various proteins." "This observation is intuitively consistent as one would expect rare SNVs to have greater impact on protein stability." "The proposed framework is a logical extension to some of the earlier studies, which primarily employed global metrics such as folding free energy changes to quantify the effects of genomic variants."

Reviewer #2

-- Ref 2.0 – Accessibility of the method --

Reviewer Comment	How can your method be accessed / used by other scientists who want to analyse their data? I don't find a link to a website / download archive or similar.
Author Response	We would first like to thank the reviewer for taking to time to carefully read through our study, as well as providing valuable suggestions on how we may improve this work. With respect to source code, we have now provided this content as a public resource on github: https://github.com/gersteinlab/Frustration

-- Ref 2.1 – Filtering datasets for comparisons--

Reviewer Comment	Concerning the datasets used for benign and disease-causing SNVs. Which variants from HGMD were included? As far as I know, there are different categories of variants in HGMD: DM=Disease causing (pathological) mutation, DM? = Likely disease causing (likely pathological) mutation, DP=Disease associated polymorphism, DFP=Disease associated polymorphism with additional supporting functional evidence, FTV=Frameshift or truncating variant with no disease association reported yet, FP=Polymorphism affecting the structure, function or expression of a gene but with no disease association reported yet. In order to create a testset of "disease mutations", all categories except for DM should be avoided in order to make sure that the test data has the highest possible quality. Variants which were found in association studies are not suitable to go into a test set of disease mutations, since there is only an association between the variant and the disease and not a proven functional link. The same applies for the data taken from 1000G and ExAC: Although these are generally denoted "common", there are significant differences in the genotype frequencies and MAFs of the variants. Especially in the ExAC data, variants which are associated with a specific clinical phenotype might be included. Moreover, there are also variants from TCGA (which went into your disease-variant set) included in ExAC.Did you chose a certain threshold for genotype frequency or MAF, above which you considered a 1000G / ExAC variant as common enough to be harmless/benign? If yes, this should go to the paper/supplement, if no, you should restrict the dataset to a somewhat smaller subset of variants, according to a sensible threshold. Moreover, did you cross-check if there are HGMD variants, which are also present in the 1000G/ExAC data? This also happened in the past.
Author Response	We thank the reviewer for these valuable suggestions. We have updated our datasets such that: 1) we only keep HGMD variants with the status label “DM”; 2) we have removed HGMD and TCGA variants present in ExAC; and 3) we have removed HGMD variants in the 1000 Genomes dataset. When this filtering is performed, a very small fraction of SNVs were removed from our analysis, and we note that this filtering did not heavily affect our main results. However, we have updated our figures and p-values to reflect this pre-processing. We describe this pre-processing scheme in the method section of the paper.

Deleted: We describe this pre-processing scheme in the Methods section of the paper.

	We applied a MAF threshold of 0.5% to distinguish between rare and common variants. This information is now provided in our updated Methods section.
Excerpt From Revised Manuscript	"In order to avoid redundancy and false positive call sets, we only consider HGMD SNVs annotated as pathological (labeled as "DM") in our HGMD dataset. Furthermore, we removed HGMD variants present in the 1000 Genomes and ExAC datasets. Similarly, we also removed known TCGA variants present in the original ExAC SNV datasets." "Furthermore, we investigated the differential influence of common and rare mutations, where SNVs with minor allele frequency (MAF) less than or equal to 0.5% were considered to be rare mutations. SNVs were otherwise classified as common."

-- Ref 2.2 – Usefulness of the method --

Reviewer Comment	To underline the usefulness of your method, which is, as said in your manuscript, to meet a "growing and urgent need to evaluate the potential effects of low-allele-frequency variants in unbiased ways using high-throughput methodologies", I miss some extra calculations / benchmarking. There are methods existing in order to evaluate potential effects of low-allele-frequency variants in unbiased ways (SIFT, PolyPhen2, MutationTaster, and many others). I would like to see how exactly your method adds up to this. Is the additional information gained from structural analysis really an advantage over existing methods? If you could show this, this would surely be an argument for people to use and cite your method. If they don't know if your method is really helpful, they will maybe not even try it, since analysis of high-throughput data is (already) time-intensive. One could for example create a small set of variants and analyse these with one or two of the "common" tools to predict the deleteriousness of SNVs (e.g. PolyPhen2 and MutationTaster2, since these are generally considered the most accurate ones) and then check if there are disease variants predicted as "harmless" by these tools (i.e. false negative) which are then correctly seen as locally maximal frustrated by your method. Or any other way how it can be shown that the method is indeed useful for the analysis of high-throughput data (e.g. compare with other existing "structural prediction" tools, if those exist).
Author Response	We are thankful to the reviewer for proposing this interesting analysis. Following the reviewer's suggestion, we ran SIFT and Polyphen2 on a smaller set of HGMD variants. These smaller set of variants were selected on the criterion that they map to 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. Frustration metric was able to rescue ~38% and ~46% of polyphen2 & SIFT annotated false negative variants, as

	described in the result and supplementary information. We also highlight few examples by plotting linear diagram for such cases in the supplementary information.
Excerpt From Revised Manuscript	*Excerpt from Results section of main text:* As discussed, existing structure-based methods for predicting SNV deleteriousness rely on global metrics of protein stability. These approaches may incorrectly predict known disease-associated SNVs to be benign (thereby producing false negatives). We address the extent to which ΔF rescues such false negatives by correctly predicting their deleterious effects. We first identified 626 HGMD SNVs within the semi-balanced set (see Method section), and predicted the impacts of these SNVs using SIFT, PolyPhen2, and ΔF values. SIFT produces false negatives for 13.7% of these HGMD SNVs. We find that ΔF rescues 46% of these SIFT false negatives (i.e., by correctly predicting deleterious impacts). Similarly, PolyPhen2 produces false negatives for 10% of the HGMD SNVs. Applying ΔF enables us to rescue 38% of these PolyPhen2 false negatives. Glucokinase is used as an example to demonstrate specific cases of rescued variants (SI Figure S7). Finally, a list of all false negatives rescued by ΔF analysis is provided in SI data file.

Deleted: We further highlight the potential complementarity of using local frustration as a means of complementing existing methods for evaluating SNV deleteriousness. These existing methods utilize global stability/conservation to predict variant deleteriousness. For this analysis, we selected a smaller set of variants mapped to PDB structures, and selected those structures such that at least one HGMD and at least one ExAC non-synonymous SNVs map. Subsequently, we identified instances in which HGMD variants were predicted to be benign by polyphen2 or SIFT (false negatives) but ΔF suggests harmful impacts. 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 not be damaging. Furthermore, we analyzed the ΔF values for variants in this dataset. Applying the ΔF threshold described earlier (-1.221), we observed that 38% of the miss-annotated variants had significantly large ΔF values, indicating their potential deleteriousness. Furthermore, we also determined that ~46% of SIFT-annotated false negative variants had large ΔF values. We also highlight an example by plotting linear diagram for such a case in the supplementary Figure S7.

Formatted: Font:9 pt

Deleted: ."

-- Ref 2.3 – Method run time scale --

Reviewer Comment	How long would it take to analyse let's say 10,000 SNVs? As this is more or less the dimension which goes along with HT-sequencing.
Author Response	The reviewer has raised a good question of practical interest. We ran our pipeline on 10,000 SNVs, and it took ~2.5 hours to map these variants to PDB structures. In total, we mapped 20% of these SNVs onto three-dimensional structures. Further, generating the mutated protein model and frustration calculations for the structurally mapped variants took ~26 hours.
Excerpt From Revised Manuscript	"This workflow is computationally tractable when evaluating ΔF for large numbers of variants. Our benchmark calculations on 10,000 non-synonymous SNVs indicates that we can map, build mutated models, and calculate ΔF values in ~29 hours on an E5-2660 v3 (2.60GHz) core."

-- Ref 2.4 –Typographical error --

Reviewer Comment	Concerning Fig. 1: Residues are not numbered. In the text, you talk about ILE in pos. 31 which is exchanged to TYR. In the figure legend, you say that TRP is changed to TYR. In the picture, there is TRP highlighted as well as TYR, but the native and mutated structure (at least the part shown) differ in more than just this one residue. This confuses me and should be clarified.
Author Response	We thank the reviewer for pointing out this inconsistency. We have fixed the text in our methods section to remove this ambiguity.
Excerpt From Revised Manuscript	"In Figure 1, we demonstrate an example case in which replacing tryptophan at a particular locus within ubiquitin (PDB ID 1UBQ) with a tyrosine."

-- Ref 2.5 – Violin plot description in figure legends --

Reviewer Comment	Concerning the Fig. 2-6 (violin plots): The figure legends do not say what the white dots and the vertical lines stand for. Mean? Median? Standard deviation? Range? This should be explained. Which difference between delta F is regarded significant (concerning differences in delta F "core" between bening SNVs and disease-causing SNVs)?
Author Response	We agree that some clarifications were needed here. In the revised manuscript, we explain the meanings of white dots and vertical lines within the updated figure legends. Comparison of ΔF distributions for the ExAC core SNVs and HGMD core SNVs point to statistically significant differences (p-value < 2e-16 using a two-sided Wilcoxon test). Furthermore, this observation was also true for comparisons involving 1000 Genomes core SNVs and HGMD core SNVs.
Excerpt From Revised Manuscript	"The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.."

-- Ref 2.6 – cutoff for common/rare differentiation --

Reviewer Comment	Fig. 3: Which MAF separates "common" from "rare" SNVs?
Author Response	We applied a MAF threshold of 0.005 to distinguish between rare (MAF <=0.005) and common variants. This previously missing information has now been incorporated into the text.
Excerpt From Revised Manuscript	"Furthermore, we investigated the differential influence of common and rare mutations, where SNVs with minor allele frequency (MAF) less than or equal to 0.5% were considered to be rare mutations. SNVs were otherwise classified as common. "

-- Ref 2.7 – Spacing error --

Reviewer Comment	Very minor point: Sometimes, spaces are missing (e.g. p.3 l.21/l.37). Re-check for this.
Author Response	We thank the reviewer for pointing this out. We have fixed this formatting error in the updated version of the manuscript.
Excerpt From Revised Manuscript	

-- Ref 3.1 – Regarding limitation of method --

Reviewer Comment	The main rationale for the paper put forward by the authors is rapidly growing number of rare variants coming from individual genomes sequencing projects and the need for new methods to infer potential functional associations of such variants. However, the results presented in this

	work clearly underscore main limitation of all structure-based methods: scarcity of high-resolution 3D protein structures and low PDB mapping coverage makes them less useful compared to more common sequence-based methods. In fact, the fraction of successfully PDB-mapped variants from ExAC database reported by the authors is below 2% (Supporting Information). This makes method's potential contribution to large scale interpretation of rare and unknown significance variants rather questionable. More general estimates usually agree upon less than 10% of all known human proteins covered by PDB, still too few. Unfortunately, there is no evidence that this coverage would increase significantly in the near future. Also, PDB is highly biased towards representing a subset of all known protein folds/domains and this bias keeps increasing, not diminishing. I would recommend either removing or significantly toning down all claims about potential applicability of the method towards large-scale human variant interpretation, specifically from the Abstract and Introduction.
Author Response	We thank the reviewer for pointing out these issues. We now discuss the limitations of this approach in order to tone down and qualify its applicability. We agree that there are inherent limitations in structure-based methods as a result of relatively low coverage across the human proteome. However, there has been a persistent increase in the structural coverage due to improvements in three-dimensional structure determination. We have highlighted this gradual increase in protein structural space in a recent review (pubmedID:26658741). In addition, we anticipate further increases in the structural coverage due to cryo-Electron microscopy. The advent of cryo-EM has made it possible to resolve the three-dimensional structures of relatively large protein/protein-complexes, which were unfathomable a decade ago. Finally, the growing systems-level view of protein biology (e.g., protein-protein interaction networks) may help to broaden the relevance of the limited number of cases in which SNVs lie within known structures (discussed in excerpt below). However, the limited coverage of SNVs in structures persists as a major challenge, so we have also provided a discussion of this challenge in the updated manuscript.
Excerpt From Revised Manuscript	*Excerpt from Introduction:* …Though the majority of disease-causing variants lie in non-coding regions of the genome, many of them lie in protein-coding genes. Furthermore, only a limited fraction of non-synonymous SNVs may be mapped to known protein structures. However, immense progress has been made in resolving the three-dimensional structure of many proteins over the last several decades (13)…. *Excerpt from Discussion:* "Historically, the relative scarcity of genomic variation and structural data have presented challenges in variant interpretation, in that only a small pool of SNVs may be mapped to resolved structures… However, limited mapping coverage persists as a major challenge, a number of recent trends may partially help to mitigate this issue. Significant improvements in crystallographic protocols have enabled near-exponential growth in deposited X-ray structures in the PDB (10). Furthermore, cryo-EM is opening entirely new avenues for revealing the architectures of many proteins which were previously elusive to crystallography, which is expected to expand the structurally-resolved proteome (59). Finally, systems-level descriptions

	of cellular phenomena provide a more complete understanding of context in which proteins operate. Specifically, there is a growing understanding of protein-protein interaction networks and the role of resolved structures therein (60). As such, inferring how a given SNV affects a particular structure is by no means limited to predictions regarding that protein alone – the protein's tight associations with other molecules may greatly broaden the scope of how that SNV influences more global cellular phenomena. For instance, the functional consequences of an SNV within a central hub protein of a network may effectively be propagated. "

-- Ref 3.2 – variant statistics and semi-balanced variants --

Reviewer Comment	Another known issue is strong annotation disparity between known Mendelian disease mutations (e.g. HGMD disease variants) and other variants: most of HGMD mutations are reported in a small subset of proteins, while majority of the proteins only have fewer and mostly benign or unknown significance variants reported for them. This creates bias when performing comparisons between the two functional classes of variants. In case of PDB-mapped variants, such annotation bias might have been alleviated to some extent by the PDB intrinsic bias (mentioned above, skews PDB & HGMD data towards the same proteins) but it requires further investigation. Authors should present statistics for the number of unique proteins and the distribution of variants in the unique proteins for each of their datasets. They should also attempt to perform their analysis on a (semi-)balanced set(s) of variants, using sets of proteins where both disease and neutral mutations are present. See Grimm et al. (2015) Human Mut. 36:513-523 for an example of such balanced sets and trends analysis.
Author Response	We thank the reviewer for these observations, and we agree that some analyses and discussion should be devoted to exploring these points. As such, new analyses and text have been integrated into the Discussion and Supplementary section of the revised manuscript. We have also performed our analysis on a semi-balanced set of variants (as proposed by the reviewer), and we report the results of this analysis in the supplementary information. Overall the trends were very much consistent with our prior analyses. However, the new dataset lacks statistical significance, potentially as a result of the fact that it is considerably smaller dataset. The details of these analyses are provided in the excerpt below.
Excerpt From Revised Manuscript	*Excerpt from Method* **Semi-balanced SNV datasets** The limited and uneven structural coverage of the human proteome primarily introduces two sources of potential bias when combined with SNV datasets: 1) some proteins may be over-represented when evaluating the effects of SNVs, and 2) the sets proteins that correspond to benign SNVs may differ considerably from those that correspond to deleterious SNVs, thereby making direct comparisons between benign and deleterious SNVs less reliable. In order to address this first issue, we select a non-redundant set of 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. Using

Formatted: Indent: First line: 0.5"

Moved (insertion) [1]

	this approach, 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 non-redundant protein sets are given in Supp. Fig. S2-S4. In order to address the second issue, we analyze only those structures that fall within the *intersection* of the different non-redundant datasets. Thus, for each SNV mapping to structure within this intersection set of non-redundant proteins (which we term the "semi-balanced set"), at least one residue overlap with an ExAC(1KG) and HGMD SNV. We utilize this semi-balanced SNV set to elucidate utility of frustration metric with respect other methods (polpyhen2 & SIFT), as described in the result section. We also perform ΔF comparison for 1KG, ExAC and HGMD variants on the semi-balanced SNV sets (Supp. Fig S5). *Excerpt from Discussion* Historically, the relative scarcity of genomic variation and structural data have presented challenges to variant interpretation, in that only a small pool of SNVs may be mapped to resolved structures. Furthermore, this limited coverage may exacerbate bias in two regards: 1) certain proteins may be over-represented in any given dataset, and 2) the proteins affected by disease-associated SNVs differ from those in which more benign SNVs intersect (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.).

Moved (insertion) [2]

Moved (insertion) [3]

Formatted: Font:Not Italic, No underline

Formatted: Line spacing: 1.5 lines

Deleted:

Formatted: Indent: First line: 0.5"

Moved up [1]: 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.

Deleted: 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.

Moved up [2]: Fig. S2-S4.

Deleted:

... [1]

Deleted:

... [2]

Moved up [3]: Fig S5).

Deleted: However, they lack statistical significance, potentially due to lower amount of SNVs included in the semi-balanced dataset."

-- Ref 3.3 – SNV frequency summary --

Reviewer Comment	Please, provide complete breakdown for the raw numbers of SNVs in each subcategory analyzed for the data presented in the Figures: Core/Surface, Core/Surface/Common/Rare, etc.
Author Response	We agree that these numbers are important to know, and indeed, reviewer #1 had the same suggestion. These statistics are now provided in Table 1, which may be found within the main text.

Excerpt From Revised Manuscript	**Table 1.** Summary statistics on the number of SNVs used in comparative analyses. Shown are variant counts for non-disease *(top)*, HGMD *(bottom-left)*, and pan-cancer SNVs *(bottom-right)*.

Conservation measure	1000 Genomes core	1000 Genomes surface	ExAC core	ExAC surface
DAF rare (common)	2267 (85)	1570 (106)	17972 (102)	11550 (83)
GERP conserved (variable)	1552 (287)	1132 (212)	12165 (2174)	7637 (1406)

Conservation measure	HGMD core	HGMD surface
GERP conserved (variable)	5158 (961)	1113 (221)

SNV type	PANCAN core	PANCAN surface
non-CAG	2153	1848
CAG	4140	2767
driver	877	486

-- Ref 3.4 – Typographical error --

Reviewer Comment	Supporting information, page 2: “SNVs are classified in three groups based Coin the native state (MinFNS)”, possibly a typing error: Coin>on? Also, item a) is missing; enumeration starts from b).
Author Response	We thank reviewer for pointing out this typographical error. We have fixed this error, and note that this paragraph has been moved to the Methods section.
Excerpt From Revised Manuscript	"SNVs are classified in three groups based on the native state a) minimally frustrated in the native state"

Page 11: [1] Deleted	Microsoft Office User	8/6/16 3:19 PM

To control for these two effects, we first identify a non-redundant set of unique proteins within each dataset.

Page 11: [2] Deleted	Microsoft Office User	8/6/16 3:19 PM

Excerpt from Supplement:
"After identifying unique protein sets, those proteins that 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.