RESPONSE LETTER

Reviewer #2

-- Ref 2.0 reporting on the ΔF threshold–--

Reviewer Comment	I suggest to add some information concerning the "deltaF-threshold" which was used to discriminate deleterious from benign (as deduced from deltaF value) variants in the SIFT/Polyphen-2 complementing analysis to the main text. Is it -1.221, as explained in supplemental information? Or any other value? This should be mentioned in the main text, otherwise the reader cannot really follow what you did. There might also be an additional methods section on this analysis in the supplement.
Author Response	We thank the reviewer for providing additional feedback on how this manuscript may be improved. In the previous version of the manuscript, we provided the ΔF threshold information in the Methods section. With respect to reporting the ΔF threshold, we now explicitly mention this cut-off value (-1.221) in the Results section of the main text of the updated manuscript. Regarding additional supplementary method section for the SIFT/Polyphen-2 complementing analysis, we already provide necessary information (selection of PDB subset for the analysis and deltaF-cutoff selection method) in the method and supplement section of the current manuscript. Thus, we think additional details will be redundant here.
Excerpt From Revised Manuscript	*Excerpt from Results:* We use a ΔF threshold of -1.221 to discriminate between SNVs that are predicted to be benign or deleterious. Details regarding how this threshold value was established are provided in the supplement.

-- Ref 2.1 – Importing a supplementary figure into the main text--

Reviewer Comment	I would also suggest to add supplemental figure S1 to the main article, since it gives a good overview of used data. Instead, Figure 2 and/or 6 could go to the supplement (if you have too many figures).
Author Response	We agree with reviewer that this figure (i.e., what was previously Figure S1) would have more value as a main text exhibit. As such, this figure now appears as Fig. 1 in the main text.

Deleted: – additional details
Deleted: deltaF-thershold --
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: y
Formatted: Left
Deleted: would first like to
Deleted: further valuable suggestions
Deleted: we may improve
Deleted: work
Deleted: "deltaF-
Deleted: "
Deleted: m
Deleted: However, following
Deleted: reviewer's suggestion
Deleted: r
Deleted: as well
Deleted: For the frustration metric, we applied
Deleted: (see method for detail)
Deleted: distinguish
Deleted: and
Deleted: variants
Deleted: Additional
Deleted: concur
Deleted: reviewer's suggestion and now include
Deleted: figure

-- Ref 2.2 – Table caption and rare/common variants--

Reviewer Comment	I am a bit surprised that there are more "rare" than "common" and more conserved than variable SNVs in the 1KG and ExAC data set(s), since intuition would tell me that it should be the other way round (since these SNVs are present in healthy human populations, and as you said, 1KG and ExAC "are highly enriched in benign SNVs".). Maybe it would help to have your definitions of rare/common (MAF?) and conserved/variable (specific GERP score?) directly in the table caption.
Author Response	The reviewer aptly points out the need for clarifications here. With respect to having more rare than common SNVs in these datasets, we would point out that we are only restricting our analyses those non-synonymous SNVs that may be mapped to protein structures. Relative to all SNVs within the genome (including synonymous SNVs, SNVs within non-coding regions, and SNVs within difficult-to-crystallize disordered protein segments), mappable non-synonymous SNVs occur at lower allele frequencies and lie within more conserved regions. Thus, the majority of the SNVs we investigate will intrinsically tend to be rare variants within conserved regions. With respect to MAF and GERP cutoffs, we have updated the caption of Table 1, which now explicitly states the MAF and GERP threshold values that are used to distinguish between rare/common and conserved/variable SNVs, respectively.
Excerpt From Revised Manuscript	**Table 1. Summary statistics on the number of SNVs used in comparative analyses.** Shown are SNV counts for non-disease (top), HGMD (bottom-left), and pan-cancer SNVs (bottom-right). Variants were further classified as being rare (MAF <= 0.5%) or common (MAF > 0.5%), as well whether or not SNVs lie within conserved (GERP > 2.0) or variable (GERP <= 2.0) genomic regions.

Deleted: We update the caption of table 1 to explicitly state the MAF and GERP cutoff values distinguishing rare/common SNVs as well as conserved/variable datasets. ... [1]

Deleted: This table shows variant

Deleted:

Deleted: %),

Deleted: %) ,

Deleted: and

Deleted:).

Deleted: F

-- Ref 2.3 –Schematic figure description--

Reviewer Comment	I also still don't get Fig.1 (although I principally like it!). According to methods text and figure capture, TRP was changed to TYR: "Shown here is the result of changing residue ID 31 in plastocyanin (pdb ID 3CVD) from the wild6type residue (TRP) to a mutated residue (TYR)". These two amino acids are also highlighted / differently colored in the figure. However, the sequence context of those two highlighted amino acids is not the same. If there were only this one amino acid exchange, shouldn't the rest of the illustrated sequence be identical? Or is the illustrated sequence of amino acids not the "real" amino acid sequence but a somehow linearized spatial configuration / structural order of the amino acids, as they appear after folding to secondary and tertiary protein structure? The figure might be easier to understand if the residues were numbered (as I suggested already before).
Author Response	Regarding Figure 1, we feel that clarifications are needed. If we understand correctly, the reviewer has interpreted the two vertical lists of amino acids as constituting a type of sequence within the protein

Deleted: We thank

Deleted: for bringing this point. Unfortunately, reviewer is confusing

Deleted: represented vertically

	(either a literal primary amino acid sequence or some other type of spatial sequence). However, these amino acids are not intended to represent any type of sequence. Each of the two vertical lines should be interpreted as energy-level diagrams. Each level on this energy scale corresponds to the total energetic value of the protein if the residue position (residue ID 31) were to be occupied by distinct amino acids (thus, for instance, if we consider the left vertical line, having isoleucine occupy position 31 results in conferring the highest possible energy to the protein, whereas having valine occupy position 31 results in the lowest possible energy for the protein). This energy is determined using an empirical energetic term, which depends on the identity of the residue and its surrounding environment. Note that we do not perform any structural modeling for this calculation. The left vertical line shows residues that are listed based on the energies that they impart in the native structure of the protein. In contrast, the right vertical line corresponds to the energies that are calculated when using the modeled protein structure (this modeled structure is one in that was built using homology modeling upon changing the TRP residue at potion 31 to TYR). In order to bring out the point regarding energetic levels more visually, we have modified the figure. We feel that some confusion may be avoided by omitting the images of protein side chains (which do indeed resemble primary sequences). In addition, we have changed the relative spacing between amino acids, such that the gaps between consecutive amino acids are no longer the same. We hope that this more clearly exhibits energetic levels, rather than sequences. We have also modified out figure caption in order to clarify this.
Excerpt From Revised Manuscript	**Figure 1: An example illustrating the case in which $\Delta F < 0$.** Each of the two vertical lines represents energy-level diagrams. Each level on this energy scale corresponds to the total energetic value of the protein if the residue position (here, residue ID 31) were to be occupied by distinct amino acids (thus, for instance, if we consider the left vertical line, having isoleucine occupy position 31 results in conferring the highest possible energy to the protein, whereas having valine occupy position 31 results in the lowest possible energy for the protein). The ΔF associated with an SNV is negative if the SNV introduces a destabilizing effect. Shown here is the result of changing residue ID 31 in plastocyanin (pdb ID 3CVD) from the wild-type residue (TRP) to a mutated residue (TYR). *Left)* The protein in its wild-type form (in green), in which the tryptophan residue at position 31 is substantially more energetically favorable relative to the mean energy $\langle E \rangle$ that would result from having any of the possible 20 amino acids at that position. This disparity is designated by $(\langle E \rangle - E_{nat})/\sigma_E = F_{nat} > 0$. *Right)* The entire protein structure is then modeled (see methods) to generate the mutated structure after the SNV W31Y is introduced, thereby changing the relative energetic distributions for the different amino acids. The new mean and standard deviation associated with the energies of the modeled structure are designated by $\langle E \rangle'$ and σ_E', respectively. In this case, the SNV W31Y results in an energy that is higher than the mean energy of all possible 20 amino acids at that position. This disparity is designated by $(\langle E \rangle' - E_{mut})/\sigma_E' = F_{mut} < 0$. Taken together, the negative value associated with the disparity between the F_{mut} and F_{nat} values ($F_{mut} - F_{nat} = \Delta F < 0$) indicates that this SNV is locally unfavorable.

Deleted: /structural context, which is incorrect. The vertical line in the schematic should be considered more like an energy level description of protein.

Deleted: y

Deleted: ,

Deleted: was

Deleted: . The total

Deleted: by

Deleted: y

Deleted: residue

Deleted: .

Deleted: represents

Deleted: their energy values

Deleted: hand

Deleted: energy level based on

Deleted: , where wild type

Deleted: was mutated

Deleted: using homology-modelling.

Deleted: employ

Deleted: modeled structure as template to further determine the energy level description of the modeled structure on the right vertical line

Deleted: .

Formatted: Heading 4

-- Ref 2.4 – Neutral terms for variants --

Reviewer Comment	I would suggest to use a neutral term for variants of not further specified clinical significance, regardless whether they are rare or common. Neutral terms are "variant", "variation", "base exchange" etc. The term "mutation" should be avoided when the clinical significance of a variant is unknown or unspecified, since it is often (mis-)understood as a variant which causes disease. Example sentence, where "mutation" should be replaced by neutral term: "Furthermore, we investigated the differential influence of common and rare mutations, where SNVs with minor allele frequencies (MAF) less than or equal to 0.5% were considered to be rare mutations."
Author Response	We agree with reviewer's suggestion for using a more neutral term in the context of variants with unknown clinical significance, and we have modified the manuscript accordingly.
Excerpt From Revised Manuscript	

Deleted: . We update

-- Ref 2.5 – Disease-associated term --

Reviewer Comment	The term "disease-associated" should be used with care in order to avoid confusion between disease-association and disease-causality. There is a dedicated method called association study, which strives to detect an association between genetic variants and a certain (mostly complex) disease, where associated variants are not necessarily causative. In contrast to this, disease-causing variants are not only statistically associated with a disease but have been shown to be causative for it, which has to be distinguished from disease-association. Therefore, some sentences should be rewritten, for example in the abstract: "disease-associated SNVs create stronger changes in localized frustration than non-disease associated variants" and in the main text "We also examined the local perturbations induced by disease-associated and benign SNVs originating in conserved and variable regions of the genome." and "[...] wherein we analyzed KF distributions for HGMD variants (disease-associated)[...]" - please check for further occurrences, also in the supplemental / supporting information. There should be clarity about the difference between disease-association and disease-causality in your manuscript. This avoids confusion on side of your readers.
Author Response	We agree with reviewer that the term "disease-associated" can be potentially confusing. In order to avoid such confusion, we now use the term "disease-related" throughout our text.
Excerpt From Revised Manuscript	"…disease-related SNVs create stronger changes in localized frustration than non-disease related variants…" "We also examined the local perturbations induced by disease-related and benign SNVs originating in conserved and variable regions of the genome."

Deleted: [[Not sure what to use instead of disease-associated.]]

-- Ref 2.6 – GERP and DAF abbreviation --

Reviewer Comment	Should be fully spelled at least once somewhere in the manuscript. Do you mean GERP = Genome Evolutionary Rate Profiling and DAF = derived allele frequency?
Author Response	We thank the reviewer for pointing out this issue. They have now been corrected.
Excerpt From Revised Manuscript	"The distinction between conserved and variable regions were defined using genome evolutionary rate profiling(GERP) scores"

-- Ref 2.7 – definition of rare/common in table caption –

Reviewer Comment	Thresholds for your definition of rare/common (MAF?) should appear in the table caption.
Author Response	We have updated the table caption to include this definition.
Excerpt From Revised Manuscript	**Table 1. Summary statistics on the number of SNVs used in comparative analyses.** This table shows variant counts for non-disease (top), HGMD (bottom-left), and pan-cancer SNVs (bottom-right). Variants were further classified as rare (MAF <= 0.5%), common (MAF > 0.5%) , conserved (GERP > 2.0) and variable (GERP <= 2.0).

Page 2: [1] Deleted	Microsoft Office User	9/17/16 3:28 PM

We update the caption of table 1 to explicitly state the MAF and GERP cutoff values distinguishing rare/common SNVs as well as conserved/variable datasets.

We would also like to point out that we are evaluating the impact of only non-synonymous SNVs in the 1KG and ExAC datasets, which map to protein structure. This primarily drives the disparity in frequency of rare/common and conserved/variable SNV datasets, which reviewer is alluding to.