Localized structural frustration for evaluating the impact of sequence variants

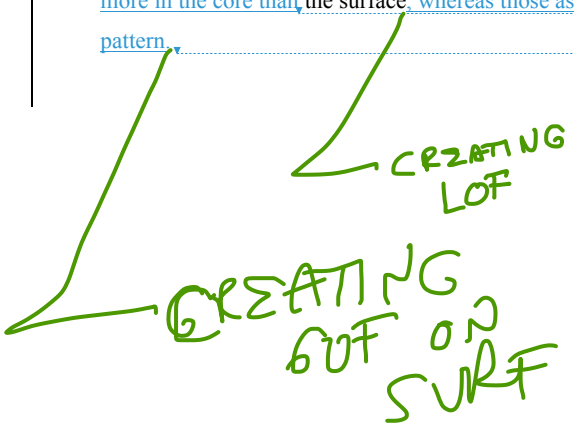
Sushant Kumar[a,b], Declan Clarke[c], Mark Gerstein[a,b,d,1]

[a]Program in Computational Biology and Bioinformatics, Yale University
[b]Department of Molecular Biophysics and Biochemistry, Yale University
[c]Department of Chemistry, Yale University
[d]Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

[1] Correspondence should be addressed to M.G. (pi@gersteinlab.org)

Abstract

Population-scale sequencing is increasingly uncovering large numbers of rare single-nucleotide variants (SNVs) in coding regions of the genome. The rarity of these variants makes it challenging to evaluate the deleteriousness of the SNVs with conventional phenotype-genotype associations. Protein structures provide a way of addressing this challenge. Previous efforts using them have focused on globally quantifying the impact of SNVs on protein stability. However, local perturbations are known to severely impact protein functionality without strongly disrupting global stability (e.g. in relation to catalysis or allostery). Here, we describe a workflow in which localized frustration, quantifying unfavorable local interactions, is employed as a metric to investigate such effects. Using this workflow on the PDB, we find that frustration produces many immediately intuitive results: for instance, disease-associated SNVs create stronger changes in localized frustration than non-disease associated variants, and rare SNVs tend to disrupt local interactions to a larger extent than common variants. Less obviously, we observe that somatic SNVs associated with oncogenes and tumor suppressor genes (TSGs) induce very different changes in frustration. In particular, those associated with TSGs change the frustration more in the core than the surface, whereas those associated with oncogenes manifest the opposite pattern.



Deleted: Growing sequence datasets are
Formatted: Font color: Auto
Deleted: is
Formatted: Font color: Auto
Deleted: renders
Formatted: Font color: Auto
Deleted: variant-
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: unfeasible when trying to evaluate the potential deleteriousness of SNVs. As such, protein
Formatted: Font color: Auto
Deleted: may help to
Formatted: Font color: Auto
Deleted: the needed means for inferring otherwise difficult-to-discern rare SNV-phenotype association.
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: sought to quantify
Formatted: Font color: Auto
Deleted: global
Formatted: Font color: Auto
Deleted:
Deleted: can
Deleted: (such as catalysis, allosteric regulation, interactions and specificity)
Deleted: .
Deleted: (which quantifies
Deleted:)
Deleted: Most of our observations are intuitively consistent: we observe that
Deleted: have a strong proclivity to induce strong
Deleted: variants
Deleted: ones. Furthermore
Deleted: stronger perturbations at
Deleted: and in the interior, respectively. These findings are consistent with the notion that gain-of-function (for
Deleted:) and loss-of-function events (for TSGs) may act through changes in regulatory interactions and basic functionality, respectively.

Introduction

The advent of next-generation sequencing technologies has led to a remarkable increase in genomic variation data at both the exome as well as the whole-genome levels (1, 2). These large datasets are playing a pivotal role in advancing efforts toward personalized medicine (3). Non-synonymous coding single nucleotide variants (termed SNVs throughout this study) are of particular interest because of their implications in the context of human health and disease (4–6). As such, considerable effort has been invested in curating disease-associated SNVs into various databases, including the Human Gene Mutation Database (HGMD) (5), ClinVar (6)and the Online Database of Mendelian Inheritance in Man (OMIM) (4). Concurrently, initiatives such as The 1000 Genomes Project (7, 8), Exome Sequencing Project (ESP) (9) and Exome Aggregation Consortium (ExAC) have generated large catalogues of SNVs within individuals of diverse phenotypes.

As the costs associated with sequencing entire human genomes and exomes continue to fall, sequencing will become routine in both medical and academic settings (10). Indeed, it may take less than a decade to reach the milestone of a million sequenced genomes (11), resulting in massive datasets of rare SNVs. This exponential growth in the number of newly discovered rare SNVs poses significant challenges in terms of variant interpretation (12). Compounding this challenge is the fact that many of these variants will be unique to single individuals. The extremely low allele frequencies of such “hyper-rare” SNVs render them too rare to draw variant-phenotype associations with confidence – unlike more common variants, the very rarity of these ultra-rare genomic signatures renders phenotypic inference through association studies extremely difficult. Together, these trends underscore a growing and urgent need to evaluate the potential effects of low-allele-frequency variants in unbiased ways using high-throughput methodologies.

Though the majority of variants lie in non-coding regions of the genome, many disease-associated variants are present 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). A large volume of high-resolution data on protein-protein, protein-ligand and protein-nucleic acid complexes is now available. This complementary evolution of

sequence and structural databases provides an ideal platform to investigate the functional and structural consequences of benign and disease-associated SNVs on protein structures. The integration of variant and structure knowledge bases will lead to a greater understanding of the biophysical mechanisms behind various diseases. In addition to gaining a better understanding of how disease-associated SNVs impart deleterious effects, this integration can be utilized to both predict the impacts of poorly understood SNVs (i.e., SNVs which are known to be deleterious, but for which a plausible biophysical or functional rationale is missing) and to prioritize SNVs based on predicted deleteriousness (14–17). We also note that this approach may aid in more intelligent and targeted design of drugs in various therapeutic contexts.

In last few decades, many studies have evaluated the impacts of SNVs by examining or predicting changes in thermodynamic stability (18–20). These approaches rely on the fact that SNVs may induce substantial changes in the folding landscape and conformational ensemble. Such changes in global stability are often quantified by calculating the folding free energy change ($\Delta\Delta G$) after mutating residues (20, 21).Importantly, however, many disease-associated SNVs introduce local structural changes without appreciably affecting folding free energy or global stability (22, 23). Such local perturbations may include disruptions in residue packing or hydrogen bond networks (24, 25) and salt bridges (26, 27). Examples of the associated effects include disruptions to catalytic centers, changes to “hotspot residues” that are responsible for interaction affinities and specificity, as well as perturbations to key allosteric sites (28–30). Changes to such residues may impart only minimal effects to the protein’s overall topology, but may nevertheless drastically influence protein behavior and functionality.

We examine the role of localized perturbations by calculating changes in the localized frustration indices(termed frustration throughout this study) (31, 32) of residues impacted by SNVs. Qualitatively, the frustration of a given residue quantifies the degree to which the residue is involved in favorable or unfavorable interactions with neighboring residues in space. The residue change that is introduced by an SNV may result in more (less) unfavorable interactions with neighboring residues, thereby increasing (decreasing) the frustration at that site. SNVs thereby act as agents that may relieve unfavorable interactions or alternatively impair local stability, depending on the nature of the amino acid substitution and the surrounding environment within the protein. Throughout this study, such changes in frustration are designated by ΔF.

The concept of frustration was originally introduced by Wolynes *et al.* to describe the protein folding landscape (31). The protein folding process is believed to follow a smooth funneled energy landscape, in which strong energetic conflicts are avoided (33–37). However, despite minimizing configurations that exhibit frustration, local frustration is essential to protein biology and function (38–40). Highly-frustrated local interactions result in micro-states of high potential energy. Such micro-states provide proteins with the avenues needed to carry out essential functions that entail a release of energy and the concomitant shifts in occupied energetic wells. Examples of processes that require these "energetic bursts" include catalysis, allosteric communication, conformational switches and proteinquakes (41), as well as protein-protein interactions (31, 42, 43). Ferriero et al. proposed a framework to compute the frustration profile of a given protein (32). The localized frustration index quantifies the contribution of each residue or residue pair in the total energy of the native structure compared to their energetic contribution in a random non-native configuration (see Methods). A native residue (residue pair) is considered to be minimally frustrated if it contains sufficient extra stabilization energy in its native state. In contrast, a sufficiently destabilizing residue (residue pair) in the protein structure is considered to be maximally frustrated (44). In addition, a residue (residue pair) is considered to be neutral when its stability profile lies between these extremes.

We take a data-driven approach to analyze ΔF profiles produced by the introduction of SNVs in a large dataset of proteins. SNVs present in healthy human populations (The 1000 Genome and ExAC projects) are highly enriched in benign SNVs. Therefore, we term SNVs in these datasets as "benign" (though we qualify this term by noting that a small subset of these SNVs may actually impart as yet undetected deleterious effects). However, within these datasets, there are various degrees along the continuum of phenotypic effects. While deleterious variants are more enriched among rare SNVs, neutral variants have stronger representation among common variants. In addition, we also quantified and compared ΔF profiles introduced by disease-associated SNVs (these SNVs were taken from the HGMD database), as well as cancer somatic variants, thereby enabling in-depth analyses of the differential effects between SNVs in driver and passenger genes.

The majority of our analyses were consistent with prior studies investigating how SNVs impact protein structures, we provide a distinct rationale through the lens of localized frustration. We observe that large disruptions in local interactions of minimally frustrated core residues

distinguishes disease-associated SNVs from benign SNVs as well as SNVs impacting driver and passenger genes in cancer. In contrast, benign SNVs in passenger genes generate larger perturbations in local interactions of minimally frustrated surface residues compared to core residues. Furthermore, comparisons between rare and common SNVs within healthy human populations indicate that rare variants induce larger disruptions in favorable local interactions compared to common variants. Moreover, we also investigated the effects of SNVs impacting conserved and variable regions of proteins, where conservation was measured across different species. For disease-associated SNVs, we detected a significant disparity between local perturbations observed due to SNVs impacting conserved regions compared to variable regions of proteins. However, no such disparity was observed for benign SNVs.

We also demonstrate how frustration may provide insights in the context of oncogenes and tumor suppressor genes (TSGs). We find that somatic SNVs in oncogenes disrupt local interactions of surface residues and potentially facilitate cancer progression through the introduction of non-specific regulatory interactions. However, SNVs in TSGs drive cancer progression through larger local perturbations in core residues. These observations indicate that SNVs intersecting TSGs and oncogenes as having loss-of-function (LOF) and gain-of-function (GOF) effects, respectively.

Methods

SNV datasets

Deleted: D

We utilized a comprehensive catalogue of non-synonymous SNVs from various resources. Our SNV dataset is divided into two broad categories (benign and disease-associated) (S1A). The benign set comprises of SNVs reported in The 1000 Genome Project (phase 3) (7) and subsets of SNVs curated from the Exome Aggregation Consortium. 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) (46), The Catalogue of Somatic Mutations in Cancer (COSMIC) (47) and the SNV dataset available from Alexandrov *et. al* (48). 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. SNVs from the pan-cancer dataset were further sub-classified (driver and passenger sets) based on whether they are mutating a driver or passenger gene. Driver genes were curated from the Vogelstein et. al. (49), where they distinguish between driver and passenger genes based on mutational patterns. They define a driver gene as an oncogene if the SNV is recurrent at the same gene loci, whereas tumor suppressor genes (TSG) are mutated throughout their length. Similarly, we sub-classified passenger genes into cancer-associated genes (CAGs) and non-cancer associated genes (non-CAGs). CAGs included genes from the cancer gene census (CGC) (50) and a curated list of 4050 genes from a previous study (51). Furthermore, we removed any driver gene present in the CAG dataset. The remaining set of genes impacted by pan-cancer SNVs constituted our non-CAG dataset.

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 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),

Moved (insertion) [1]

Moved (insertion) [2]

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).

Workflow to calculate frustration

As mentioned earlier, we investigated the impact of different categories of SNVs on the local stability of various protein structures. We utilize the ΔF values of mutated residues to quantify SNVs induced local perturbation. Quantifying ΔF involves three steps: a) mapping SNVs onto the affected three-dimensional structure, b) generating the homology model of the mutated structure, and c) evaluating the ΔF of mutated residue in the native and mutated conformations.

To map SNVs onto protein structures, the Variant Annotation Tool (VAT) (52) was applied to annotate our curated catalogue of SNVs. This annotation includes the gene and transcript names, residue position in the protein sequence, as well as the original and mutated residue identity. 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. We restricted this SNV mapping scheme to high-quality structures with resolution values that were better than 2.0 Angstrom. Following the SNV mapping to PDB structures, we generated models of the resultant mutated structures by applying homology modeling using the mutated protein sequence and native protein structure as input to modeler (54, 55).

Finally, we quantify the frustration index of the mapped residue in the native structure as well as in the mutated model of the protein. Briefly, the residue level localized frustration index (44) quantifies the degree to which that amino acid favorably contributes to the energy of the system relative to all 20 possible amino acids at that position:

$$F_i = \frac{<E_i^{T,U}> - E_i^{T,N}}{\sqrt{1/N \sum_{k=1}^{n} (E_i^{T,U} - <E_i^{T,U}>)}}$$, where $E_i^{T,N}$ is the total energy of the protein in the native state. The total native energy is calculated using a function that includes an explicit water interaction term, $E_i^{T,N} = \sum_{k \neq i}^{n} (E_{contact}^{i;k} + E_{water}^{i;k}) + E_{burial}^{i}$. This function, termed the associated water-mediated (AWM) potential [44], describes the energies associated with direct interactions between residues *i* and *k* ($E_{contact}^{i;k}$) as well as those with water-mediated interactions between residues *i* and *k* ($E_{water}^{i;k}$) and energy term (E_{burial}^{i}) associated with the burial of the residue. The average energy of the decoy conformations ($< E_i^{T,U} >$) is generated by mutating

Deleted: b

the original residue *i* to each of the alternative possible nineteen residues. The AMW potential includes different parameter values for different residues, so the decoy energies calculated vary based on the identity of the mutated residue. 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.

In Figure 1, we demonstrate an example case in which replacing tryptophan at a particular locus within ubiquitin (PDB ID 1UBQ) with a tyrosine (shown on left, in green). For the native structure of this protein, 19 decoy energies are calculated by changing the parameter values that are specific to each amino acid within the potential function (note that the structure is not altered or minimized in any way). In this case, the energy computed using the wild-type residue (TRP) is substantially lower than the mean value (rendering ΔE_{nat} greater than 0). Because ΔE_{nat} is greater than 0, this wild-type tyrosine is said to be "minimally frustrated".

This same protein is known to contain a disease-associated SNV at locus 31. Specifically, the disease-associated change occurs when the tryptophan is mutated to tyrosine. To quantify ΔF in this example, we first introduce the tyrosine at locus 31 *in silico*, and then use Modeler to generate a model of the mutated structure (shown at right, in orange). Thus, we now not only change the type of residue at locus 31, but also the configuration of the entire protein; the structure is said to be "non-native" (the relative energy values given on the horizontal axis may thus become redistributed slightly). In this new energy landscape, the energy associated with the residue at the mutated locus 31 is higher than the mean energy among all 20 amino acids within the modeled structure ($\Delta E_{mut} < 0$), suggesting that the mutated residue is "maximally frustrated". We are primarily interested in the ΔF between these two states. This value is proportional to the difference between ΔE_{mut} and ΔE_{nat}. ($\Delta E_{mut} - \Delta E_{nat} = \Delta\Delta E$) Here, ΔΔE is less than 0, suggesting that the frustration is higher in the mutated structure than that of the wild type.

Deleted: .
Deleted: S
Deleted: the
Deleted: (
Deleted:) is
Deleted: (i.e., wild-type
Deleted:). The vertical axis designates the different energies that would result when the residue at
Deleted: locus is mutated to each one of the other 19 amino acids. Specifically, these
Deleted: only
Deleted: In the sense that these energies are calculated in the context of the structure that is otherwise identical to the wild-type X-ray structure, the energy distribution shown at left represents the energies in "native structure". The dotted line represents the mean value among all of the 20 energy values associated with the various amino acids.
Deleted: is
Deleted: that

Downstream Analyses

In order to investigate the differential effect of SNVs in various datasets, we 'bin' each SNV into distinct categories based on their frustration index and relative accessible surface area (RSASA) in the native structure. SNVs are classified in three groups based on the native state a) minimally frustrated in the native state (MinFNS); b) maximally frustrated in the native state (MaxFNS)

and c) neutral in the native state (NeutFNS). MinFNS residues have frustration indices greater than or equal to 0.78, whereas MaxFNS residues have frustration less than or equal to -1.0. Residues falling in between these two extremes are considered in the NeutFNS category. 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 (56). Residues were defined to be in the core if their the RSASA value was less than or equal to 25 % and surface residues had RSASA values greater than 25%.

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. Similarly, we also compared the effect of SNVs influencing the conserved region and variable regions of the genome. The distinction between conserved and variable regions of the genome were defined suing GERP scores (57). GERP score identifies functionally constrained genomic elements based on multiple sequence alignment of genomic sequences from diverse species. In our analysis, we defined a genomic position as conserved if its GERP score was greater than 2.0 (GERP > 2.0). We considered genomic location to be variable, when the GERP score was positive and less than or equal to 2.0(GERP <= 2.0).

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.

Results

Differential effects of benign and disease-associated SNVs on ΔF profiles

We performed a comparative analysis to investigate the impacts of benign (1KG & ExAC) and disease-associated (HGMD) SNVs on the ΔF profiles of mutated residues in a large number of proteins. As detailed in Methods, each SNV dataset was divided into three distinct categories based on the frustration index of the wild-type residue. Maximally frustrated residues in the native structure exhibit conflicting interactions and unfavorable geometry in their local environment, thereby inducing local destabilization. Conversely, minimally frustrated residues are involved in biophysically favorable local interactions, and thus favorably contribute to the protein's stability.

For each SNV, ΔF was calculated as follows (Figure 1). For a given SNV mapped to a PDB structure, two protein structures are used in our analysis: the native structure (as it exists in the PDB), and a model of the structure as it may exist when the affected residue is mutated (this is modeled by optimizing the structure after introducing the SNV). If a given SNV maps to residue location j within the structure, then within each of these two structures, the frustration index is calculated at residue j (the corresponding values are denoted as F_{nat} and F_{mut} for the native and mutated model structures, respectively). Subsequently, we determine the difference between the frustration index of the wild-type residue in the native structure and the mutated residue in the modeled structure ($\Delta F = F_{mut} - F_{nat}$).

After calculating the ΔF values in all three categories, the resultant distributions are plotted (further details are given under Methods). We observed that most SNVs (across all datasets) affecting maximally frustrated residues in the native structure induce small but positive ΔF values. This suggests that changes to maximally frustrated residues alleviate conflicting interactions, thereby resulting in a positive frustration difference ($\Delta F > 0$). In contrast (and as expected), residues that are originally minimally frustrated tend to become more frustrated upon mutation, thereby, leading to a negative frustration difference ($\Delta F < 0$) in majority of cases across each dataset. However, we emphasize that losses or gains in favorable interactions are dependent on the type of SNV (benign or disease-associated) as well as whether the SNV affecting a surface or core residue.

We observed that benign SNVs lead to greater disruptions within minimally frustrated surface residues compared to core residues in the native structure, and this trend is observed when using both ExAC and 1KG datasets (*p-value < 2e-16 from two-sample Wilcoxon test*) (Figure 2A & 2B). In addition, disease-associated SNVs (from HGMD) result in similar

frustration changes between core and surface residues. 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*) (Figure 2C).

Differential effects of rare and common SNVs on localized frustration

In population-level studies, SNVs with lower minor allele frequencies (MAF) are generally interpreted as being more likely to be deleterious than SNVs with higher MAF values. Thus, within the set of benign SNVs provided in the 1000 Genomes and ExAC SNVs, MAF may be used as an approximation for varying degrees of selective constraint. This prompted us to compare the rare and common SNVs induced ΔF distribution for minimally frustrated core and surface residues. Consistent with our earlier observations regarding benign SNVs, we found larger disruptions to favorable local interactions in surface residues relative to core residues (Figure 3A). However, this disparity was slightly more pronounced for rare SNVs compared to common SNVs. This observation was consistent for the 1000 Genome (Figure 3A) and ExAC datasets (Figure 3B) (*with p-value < 2e-16 from two-sample Wilcoxon test*). Furthermore, using both of these datasets, we observed that greater ΔF associated with the introduction of SNVs (in either the positive or negative directions) tend to be associated with lower MAF values (Figures 3C, top & bottom panels). This trend is observed for SNVs that occur on both the surface and within the core.

Differential effects of benign and disease-associated SNVs in different evolutionary contexts

We also examined the local perturbations induced by disease-associated and benign SNVs originating in conserved and variable regions of the genome. We plotted distributions for the ΔF values for the surface and core residues (Figure 4). We observed that benign SNVs originating in both conserved and variable regions of the genome had similar effects on minimally frustrated core residues (Figure 4A & 4B). This observation was true for the surface residues as well. In contrast, disease-associated SNVs intersecting with conserved and variable genomic regions lead to variable ΔF values for surface residues (*p-value = 0.004715 from two-sample Wilcoxon test*). This disparity is even more pronounced in core residues (*p-value = 6.723e-09 from two-sample Wilcoxon test*) (Figure 4C).

Differential effects of SNVs on driver and passenger genes

One of the most important challenges confronting the cancer genomics community involves discriminating between highly deleterious driver SNVs and the large number of neutral passenger SNVs that naturally arise over the course of tumor progression (58). As part of these efforts, a large number of cancer actionable genes have been curated in recent years. We applied our framework to evaluate the effects that somatic cancer SNVs have on driver genes (49), cancer-associated genes (CAGs) (51), and non-cancer associated genes (non-CAGs) in the context of frustration. We mapped the somatic pan-cancer SNVs that intersect these three distinct gene categories onto protein structures. We then evaluated the ΔF distributions in all three categories.

As with benign SNVs, we observed that somatic SNVs impacting CAGs and non-CAGs lead to greater disruptions in minimally frustrated surface residues relative to core residues (*p-value < 2.2e-16 from two-sample Wilcoxon test*) (Figure 5). Moreover, this variability in ΔF distributions between core and surface residues was more pronounced among non-CAGs compared to CAGs (Figure 5). In contrast, SNVs that impact driver genes lead to larger disruptions in favorable localized interactions for surface and core residues (*p-value < 2.2e-16 from two-sample Wilcoxon test*) (Figure 5) compared to CAG core and surface residues.

Differential effects of SNVs on oncogenes and tumor-suppressor genes

Cancer driver genes are classified as oncogenes and tumor suppressor genes based on their mutational pattern and their mode of inducing tumorigenesis (49). Oncogenes are marked by recurrent SNVs within the same gene loci across different cancer types, and are believed to drive cancer progression through gain-of-function (GOF) mechanisms. In contrast, a tumor suppressor gene generally contains protein-truncating mutations or SNVs that are scattered throughout the gene, and they are believed to facilitate cancer progression through loss-of-function (LOF) mechanisms. This line of thinking is guided by the idea that LOF variants often act by destabilizing the protein (Figure 6C, left panel), whereas GOF variants may impact protein-protein interaction interfaces (by reducing specificity for binding partners) or negatively affect auto-regulatory sites on the protein, many of which are on the surface (Figure 6C, right panel).

In order to evaluate the extent to which such effects manifest in our set of tumor-suppressor genes and oncogenes, we applied the frustration framework to evaluate changes in local perturbation when SNVs impact these distinct categories of driver genes (Figure 6A& 6B). We observed that SNVs affecting TSGs induce stronger perturbations in minimally frustrated core residues relative to surface residues (*p-value = 0.004765 from two-sample KS test*) (Figure 6A). In contrast, SNV affecting oncogenes induces greater ΔF values within minimally frustrated residues in the surface relative to core residues (*p-value = 1.91e-13 from two-sample KS test*) (Figure 6B). Moreover, SNVs impacting *oncogenes* lead to larger disruptions in favorable local interactions compared to *TSGs* for minimally frustrated surface residues (*p-value = 2.3e-3 from two-sample KS test)*. However, SNVs impacting TSGs lead to greater disruption in favorable local interactions compared to oncogenes affecting driver SNVs in core residues (*p-value = 6.753e-13 from two-sample KS test*).

Formatted: Font:Font color: Auto

Formatted: Font:Font color: Auto

Formatted: Font:Font color: Auto

Formatted: Font:Font color: Auto

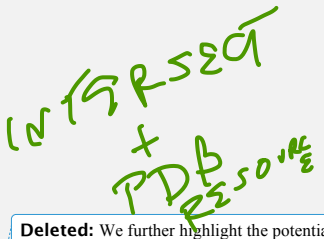
Formatted: Font:Font color: Auto

Formatted: Font:Font color: Auto

Localized frustration as a means of complementing global metrics

Deleted: to

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 provide list of these rescued false negative variants as supplement data. In addition, we also highlight an example by plotting linear diagram for such a case in the supplementary Figure S7.

Discussion

Over the course of 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 concomitant growth in the sequence and structural space provide us

with an ideal platform to investigate the impact of genomic variants on protein structure. The objective of these studies is to gain mechanistic insights into the origin of various diseases, as well as design effective drug targets for them. Prior studies in this direction were limited due to lack of genomic variation and structural data. Moreover, these studies primarily focused on investigating the impact of SNVs on the *global* stability of protein structure. However, many experimental studies have clearly indicated a causal role for SNVs inducing a local perturbation in various diseases. In this work, we repurpose the concept of localized frustration, originally introduced in protein folding studies to quantify SNV-induced local perturbations. The frustration index of a residue quantifies the presence of favorable/dis-favorable local interactions in the protein structure compared to a random molten globule structure.

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.

However, despite addressing sources of bias, limited mapping coverage persists as a major challenge. Nevertheless, 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, 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 other proteins. For instance, the functional consequences of an SNV within a multi-protein complex may adversely affect all members of the complex, despite the fact that the genomic coordinate of the SNV maps to one protein.

Deleted: .).

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

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.

In this study, we employed an extensive catalogue of benign (~5.7 million) and disease-associated (~0.76 million) SNVs. The benign SNV dataset comprised of SNVs from the 1000 Genome project (phase 3) and the ExAC project. In contrast, HGMD SNVs and pan-cancer somatic SNVs constituted our disease-associated SNV dataset. We mapped ~0.2 million benign and disease-associated SNVs onto ~10K high-resolution protein structures. Subsequently, we compared the impact of benign and disease SNVs on the frustration profile of minimally frustrated residues in various protein structures. The ΔF distributions indicated that both benign and disease SNVs disrupt minimally frustrated surface residues to similar extents. However, the mechanistic difference between benign and disease SNVs can be attributed to their impact on the local environment of core residues. Within the core, disease-associated SNVs result in more severe perturbations to local interactions relative to those introduced by benign SNVs. These local disruptions are propagated throughout the core and, in turn, drive the deleteriousness of various disease-associated SNVs.

Furthermore, we quantified the influence of rare and common SNVs present in healthy human population on the frustration profile of affected protein residues. We observed that rare SNVs lead to larger local perturbation of minimally frustrated surface residues compared to common SNVs. This observation is intuitively consistent as one would expect rare SNVs to have greater impact on protein stability. In addition, we also investigated the differential impact of SNVs intersecting conserved regions compared to variable regions of the genome. The distinction between conserved and variable regions of the genome was based on GERP scores, which quantifies a cross-species conservation score on each nucleotide position of the genome. This cross-species conservation analysis indicated that there is no disparity between ΔF associated with benign SNVs fixated in conserved and variable regions. This lack of disparity can be attributed to the absence of significant local perturbations induced by benign SNVs, which do not compromise the overall stability of protein structure. In contrast, for disease SNVs originating in conserved and variable regions of the genome, we observe significant differences in ΔF values. This is consistent with prior studies, which indicate that the deleteriousness of an SNV is more pronounced when SNVs impact functionally important conserved regions of the genome compared to variable regions of the genome.

In addition to studying disease variant in general, tremendous progress in next generation sequencing has lead to unprecedented efforts to characterize cancer genomes. Large efforts have

been invested in discriminating between driver and passenger SNVs. Driver SNVs are known to play important roles in driving cancer progression. Motivated by this, we examined the influence of SNVs emanating in driver and passenger genes. Specifically, we studied these effects in the context of the local stability of protein structure. Our analysis indicated that SNVs influencing non-actionable genes (non-CAGs) and indirectly actionable genes (CAGs) lead to greater perturbations of surface residues compared to core residues. In contrast, SNVs that impact driver genes have similar affects on ΔF values in core and surface residues. These observations further reiterate our earlier conclusion that the deleteriousness of a given SNV is determined by its ability to perturb the local interactions of core residues. These local perturbations further propagate through the core to completely destabilize the protein structure.

Furthermore, cancer driver genes are often classified as oncogenes and tumor suppressor genes based on their mode of cancer progression. SNVs in oncogenes lead to cancer progression through GOF mechanism, whereas SNVs impacting tumor suppressor genes contribute to cancer growth through LOF events. These two distinct mode prompted us to closely inspect SNVs originating in oncogenes and TSGs. We compared the ΔF profile for residues influenced by these two distinct categories of SNVs. we observed that SNVs in oncogenes and TSGs generate greater ΔF values for surface and core, respectively.

Comprehensive catalogues of genomic variations from large-scale genomics projects have clearly established the important roles of disease-associated and rare variants in human populations. We foresee further growth in genomic datasets as large-scale genomic consortia (such as International Cancer Genomics Consortium, The Pan-Cancer Genome Atlas, The UK10K Project and Mendelian Genomic Program) continue to decipher the mutational landscape of human genomes and exomes. Similarly, advances in electron microscopy, NMR, small angle X-ray scattering and other techniques will further increase the availability of protein structural data. These expanding knowledge bases of genomic variation and structural biology will facilitate integrative studies to gain mechanistic insights into disease progression and to design effective disease therapy regimens. In this work, we demonstrate the role of localized frustration as a metric to quantify and investigate the influence of genomic variants on protein structures. 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. We believe that the combination of these global and local metrics, along with

sequence features, will help to elucidate the mechanisms as well as predict the impacts of genomic variations.

Acknowledgments

We acknowledge support from the NIH and from the AL Williams Professorship funds. We thank Diego Ferreiro for helpful discussion and sharing the original source code for localized frustration calculations. We also acknowledge help of Anurag Sethi and Suganthi Balasubramanian for providing valuable feedbacks for improving the manuscript. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about

References

1. Muir,P., Li,S., Lou,S., Wang,D., Spakowicz,D.J., Salichos,L., Zhang,J., Weinstock,G.M., Isaacs,F., Rozowsky,J., *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.*, **17**, 53.
2. Soon,W.W., Hariharan,M., Snyder,M.P., Abdulla,M., Ahmed,I., Assawamakin,A., Bhak,J., Brahmachari,S., Calacal,G., Chaurasia,A., *et al.* (2014) High-throughput sequencing for biology and medicine. *Mol. Syst. Biol.*, **9**, 640–640.
3. Chen,R., Mias,G.I., Li-Pook-Than,J., Jiang,L., Lam,H.Y.K., Chen,R., Miriami,E., Karczewski,K.J., Hariharan,M., Dewey,F.E., *et al.* (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, **148**, 1293–1307.
4. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**.
5. Stenson,P.D., Mort,M., Ball,E. V., Shaw,K., Phillips,A.D. and Cooper,D.N. (2014) The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical

and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

6. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**.
7. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
8. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
9. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G., *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–9.
10. Sethi,A., Clarke,D., Chen,J., Kumar,S., Galeev,T.R., Regan,L. and Gerstein,M. (2015) Reads meet rotamers: Structural biology in the age of deep sequencing. *Curr. Opin. Struct. Biol.*, **35**, 125–134.
11. Collins,F.S. and Varmus,H. (2015) A New Initiative on Precision Medicine. *N. Engl. J. Med.*, **372**, 793–5.
12. Zuk,O., Schaffner,S.F., Samocha,K., Do,R., Hechter,E., Kathiresan,S., Daly,M.J., Neale,B.M., Sunyaev,S.R. and Lander,E.S. (2014) Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E455–64.
13. Rose,P.W., Prlić,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J., *et al.* (2014) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–56.
14. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
15. Adzhubei,I.A. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–9.
16. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, 10.1002/0471142905.hg0720s76.

17. Wong,W.C., Kim,D., Carter,H., Diekhans,M., Ryan,M.C. and Karchin,R. (2011) CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, **27**, 2147–2148.
18. Zhang,Z., Wang,L., Gao,Y., Zhang,J., Zhenirovskyy,M. and Alexov,E. (2012) Predicting folding free energy changes upon single point mutations. *Bioinformatics*, **28**, 664–71.
19. Stefl,S., Nishi,H., Petukh,M., Panchenko,A.R. and Alexov,E. (2013) Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.*, **425**, 3919–3936.
20. Kellogg,E.H., Leaver-Fay,A. and Baker,D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.*, **79**, 830–838.
21. Benedix,A., Becker,C.M., de Groot,B.L., Caflisch,A. and Böckmann,R. a (2009) Predicting free energy changes using structural ensembles. *Nat. Methods*, **6**, 3–4.
22. Lori,C., Pasquo,A., Montanari,R., Capelli,D., Consalvi,V., Chiaraluce,R., Cervoni,L., Loiodice,F., Laghezza,A., Aschi,M., *et al.* (2014) Structural basis of the transactivation deficiency of the human PPAR?? F360L mutant associated with familial partial lipodystrophy. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **70**, 1965–1976.
23. Monticone,S., Bandulik,S., Stindl,J., Zilbermint,M., Dedov,I., Mulatero,P., Allgaeuer,M., Lee,C.C.R., Stratakis,C.A., Williams,T.A., *et al.* (2015) A case of severe hyperaldosteronism caused by a de novo mutation affecting a critical salt bridge Kir3.4 residue. *J. Clin. Endocrinol. Metab.*, **100**, E114–E118.
24. Doss,C.G.P. and NagaSundaram,N. (2012) Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: A molecular dynamics approach. *PLoS One*, **7**.
25. Kumar,A., Rajendran,V., Sethumadhavan,R. and Purohit,R. (2013) Molecular Dynamic Simulation Reveals Damaging Impact of RAC1 F28L Mutation in the Switch I Region. *PLoS One*, **8**.
26. Boccuto,L., Aoki,K., Flanagan-Steet,H., Chen,C.F., Fan,X., Bartel,F., Petukh,M., Pittman,A., Saul,R., Chaubey,A., *et al.* (2014) A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. *Hum. Mol. Genet.*, **23**, 418–433.
27. Zhang,Z. (2013) A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. *Hum. Mol. Genet.*

28. Tsai, C.-J. & Nussinov,R. (2014) The free energy landscape in translational science: how can somatic mutations result in constitutive oncogenic activation? *PCCP*.
29. Li,M., Petukh,M., Alexov,E. and Panchenko,A.R. (2014) Predicting the impact of missense mutations on protein-protein binding affinity. *J. Chem. Theory Comput.*, **10**, 1770–1780.
30. Clarke,D., Sethi,A., Li,S., Kumar,S., Chang,R.W.F., Chen,J. and Gerstein,M. (2016) Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure*, 10.1016/j.str.2016.03.008.
31. Ferreiro,D.U., Hegler,J.A., Komives,E.A. and Wolynes,P.G. (2007) Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 19819–24.
32. Jenik,M., Parra,R.G., Radusky,L.G., Turjanski,A., Wolynes,P.G. and Ferreiro,D.U. (2012) Protein frustratometer: A tool to localize energetic frustration in protein molecules. *Nucleic Acids Res.*, **40**.
33. Onuchic,J.N., Luthey-Schulten,Z. and Wolynes,P.G. (1997) Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, **48**, 545–600.
34. Chavez,L.L., Onuchic,J.N. and Clementi,C. (2004) Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.*, **126**, 8426–8432.
35. Clementi,C. and Plotkin,S.S. (2004) The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.*, **13**, 1750–1766.
36. Koga,N. and Takada,S. (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.*, **313**, 171–80.
37. Frauenfelder,H., Sligar,S. and Wolynes,P. (1991) The energy landscapes and motions of proteins. *Science (80-.).*, **254**, 1598–1603.
38. Camilloni,C. and Sutto,L. (2009) Lymphotactin: how a protein can adopt two folds. *J. Chem. Phys.*, **131**, 245105.
39. Ferreiro,D.U., Hegler,J.A., Komives,E.A. and Wolynes,P.G. (2011) On the role of frustration in the energy landscapes of allosteric proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 3499–503.
40. Yang,S., Cho,S.S., Levy,Y., Cheung,M.S., Levine,H., Wolynes,P.G. and Onuchic,J.N. (2004) Domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 13786–13791.

41. Miyashita,O., Onuchic,J.N. and Wolynes,P.G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 12570–5.
42. Changeux,J.-P. (2013) 50 Years of Allosteric Interactions: the Twists and Turns of the Models. *Nat. Rev. Mol. Cell Biol.*, **14**, 819–29.
43. Zhuravlev,P.I. and Papoian,G. a (2010) Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework.
44. Ferreiro,D.U., Komives,E. a. and Wolynes,P.G. (2013) Frustration in Biomolecules. *Q. Rev. Biophys.*, **47**, 1–97.
45. Davoli,T., Xu,A.W., Mengwasser,K.E., Sack,L.M., Yoon,J.C., Park,P.J. and Elledge,S.J. (2013) XCumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**.
46. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–20.
47. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S., *et al.* (2015) COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
48. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S. a J.R., Behjati,S., Biankin,A. V, Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–21.
49. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz Jr.,L.A. and Kinzler,K.W. (2013) Cancer Genome Landscapes. *Science (80-.).*, **339**, 1546–1558.
50. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
51. Cheng,F., Jia,P., Wang,Q., Lin,C.C., Li,W.H. and Zhao,Z. (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol. Biol. Evol.*, **31**, 2156–2169.
52. Habegger,L., Balasubramanian,S., Chen,D.Z., Khurana,E., Sboner,A., Harmanci,A., Rozowsky,J., Clarke,D., Snyder,M. and Gerstein,M. (2012) Vat: A computational framework to functionally annotate variants in personal genomes within a cloud-computing

environment. *Bioinformatics*, **28**, 2267–2269.

53. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G., *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–98.
54. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
55. Webb,B. and Sali,A. (2014) Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics*, **47**, 5.6.1–32.
56. Hubbard,S.J., Campbell,S.F. and Thornton,J.M. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, **220**, 507–530.
57. Cooper,G.M., Stone,E.A., Asimenos,G., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
58. Ding,L., Wendl,M.C., McMichael,J.F. and Raphael,B.J. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.*, **15**, 556–570.
59. Bai,X., McMullan,G., Scheres,S.H.W., Ruska,H., Ruska,H., al., et, Brenner,S., Horne,R.W., Henderson,R., Unwin,P.N., *et al.* (2015) How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.*, **40**, 49–57.
60. Clarke,D., Bhardwaj,N. and Gerstein,M.B. (2012) Novel insights through the integration of structural and functional genomics data with protein networks. *J. Struct. Biol.*, **179**, 320–326.

Figure Legends

Figure 1: An example illustrating the case in which $\Delta F < 0$. 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 that introduces 31Y 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 the this SNV is locally unfavorable.

Figure 2: Differential effects of "benign" and disease-associated SNVs on the localized frustration of minimally frustrated residues in the non-mutated (i.e., native) state. Violin plots showing ΔF distributions associated with SNVs affecting the core or surface, with SNVs taken from *A)* 1000 Genomes, *B)* ExAC and *C)* HGMD. Comparison between ΔF distributions for core and surface residues of the 1000 Genomes and ExAC datasets indicate that favorable interactions of surface residues in the native states are highly disrupted upon mutation compared to core residues. Furthermore, ΔF in HGMD core residues were highly negative compared to 1KG and ExAC variants impacting core residues. The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.

Figure 3: Common and rare SNVs differentially influence ΔF profiles of minimally frustrated core and surface residues. Violin plots show ΔF distributions induced by common and rare variants present in the 1000 Genomes *(A)* and ExAC *(B)* datasets (shown are the effects on minimally frustrated core and surface residues). Rare variants in both datasets lead to more substantially negative ΔF values compared to common variants. *C)* Scatter plots of ΔF values

indicate that more extreme ΔF values (in either the positive or negative direction) tend to be associated with lower-MAF (i.e., rare SNVs). The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.

Figure 4: Comparisons between ΔF distributions associated with "benign" and disease-associated variants on evolutionarily conserved and variable residues. Violin plots depicting ΔF distributions introduced by *A)* 1000 Genomes, *B)* ExAC and *C)* HGMD variants, respectively. ΔF distributions associated with HGMD SNVs indicate larger disruption of conserved core residues compared to variable residues. In contrast, for the 1000 Genomes and ExAC datasets, no significant difference in ΔF distributions was observed for conserved and variable core residues (the same was true for surface residues). The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.

Figure 5: Comparisons between ΔF distributions associated with driver and passenger genes. *Left)* Violin plots showing ΔF distributions associated with somatic SNVs affecting *non-cancer associated genes (non-CAG), cancer associated genes (CAG) and driver genes* encoding core and surface residues. Somatic SNVs affecting core residues of driver genes lead to a more substantially negative ΔF values compared to those in CAG and non-CAG proteins. *Right)* On the contrary, SNVs in CAGs and non-CAGs disrupt favorable interactions of the surface residues to a larger extent compared to their core residues. The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.

Figure 6: Differential impacts on ΔF distributions associated with of SNVs on driver and passenger genes. Violin plots demonstrating ΔF distributions associated with SNVs tumor suppressors *(A)* and oncogenes *(B)*. SNVs in tumor suppressors (TSGs) lead to larger disruptions for minimally frustrated core compared to surface residues. However, SNVs affecting oncogenes

are associated with larger ΔF values for the surface compared to core residues. *C)* These observations suggest a potential model in which SNVs in TSGs act by disrupting the hydrophobic core of a protein and drive cancer progression through LOF mechanisms *(left)*. In contrast, SNVs in oncogenes may facilitate non-specific binding by changing surface residues and drive cancer through GOF events *(right)*. The white dots, the black boxes and vertical lines represents the medians, interquartile ranges, and 95% confidence intervals of ΔF distributions, respectively.