Localized structural frustration for evaluating the impact of sequence variants

Sushant Kumar[1,2], Declan Clarke[1,2,3] and Mark Gerstein[1,2,4,*]

[1]Program in Computational Biology and Bioinformatics, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA, [2]Department of Molecular Biophysics and Biochemistry, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA, [3]Department of Chemistry, Yale University, 225 Prospect Street, New Haven, CT 06520, USA and [4]Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

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 their deleteriousness with conventional phenotype–genotype associations. Protein structures provide a way of addressing this challenge. Previous efforts have focused on globally quantifying the impact of SNVs on protein stability. However, local perturbations may 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 Protein Databank, we find that frustration produces many immediately intuitive results: for instance, disease-related SNVs create stronger changes in localized frustration than non-disease related 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 (by introducing loss-of-function events), whereas those associated with oncogenes manifest the opposite pattern, creating gain-of-function events.

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 data sets 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-related SNVs into various databases, including the Human Gene Mutation Database (HGMD) (5), ClinVar (6) and the Online Database of Mendelian Inheritance in Man (4). Concurrently, initiatives such as The 1000 Genomes Project (7,8), Exome Sequencing Project (9) and Exome Aggregation Consortium (ExAC) (10) have generated large catalogues of SNVs within individuals of diverse phenotypes in general.

As the costs associated with sequencing human genomes and exomes fall, sequencing will become routine in both medical and academic settings (11). Indeed, it may take less than a decade to reach the milestone of a million sequenced genomes (12), resulting in massive data sets of rare SNVs. This exponential growth in the number of newly discovered rare SNVs poses significant challenges in terms of variant interpretation (13). 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-related 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 structures of many

---

proteins over the last several decades (14). In addition, 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-related 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-related SNVs impart deleterious effects, this integration can be utilized to both predict the impacts of poorly understood SNVs (i.e. SNVs that are known to be deleterious, but for which a plausible biophysical or functional rationale is missing) and to prioritize SNVs based on predicted deleteriousness (15–18). We also note that this approach may aid in more intelligent and targeted design of drugs in various therapeutic contexts.

Over the last several decades, many studies have evaluated the impacts of SNVs by examining or predicting changes in thermodynamic stability (19–21). 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 (21,22). Importantly, however, many disease-related SNVs introduce local structural changes without appreciably affecting folding free energy or global stability (23,24). Such local perturbations may include disruptions in residue packing or hydrogen bond networks (25,26) and salt bridges (27,28). 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 (29–31). 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 (32,33) 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 (32). The protein folding process is believed to follow a smooth-funneled energy landscape, in which strong energetic conflicts are avoided (34–38). However, despite minimizing configurations that exhibit frustration, local frustration is essential to protein biology and function (39–41). 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 (42), as well as protein–protein interactions (32,43,44).

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 to the total energy of the native structure compared to their contribution in a random non-native configuration (see Materials and 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 (45). 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 from SNVs in a large data set of proteins. SNVs in healthy human populations (The 1000 Genomes and ExAC projects) are highly enriched in benign SNVs. Therefore, SNVs in these data sets are termed '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 data sets, there are various degrees along the continuum of phenotypic effects. While deleterious variants are more enriched among rare SNVs, neutral variants have greater representation among common variants. In addition, we also quantified and compared ΔF profiles introduced by disease-related SNVs (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.

Though the majority of our analyses are 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-related 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-related 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 suggest that SNVs in TSGs and oncogenes may impart loss-of-function (TSG) and gain-of-function (GOF) effects, respectively.

MATERIALS AND METHODS

SNV data sets

We utilized a comprehensive catalogue of non-synonymous SNVs from various resources. Our SNV data set is divided into two broad categories (benign and disease-related; Figure 1A). The benign set comprises SNVs reported in The 1000 Genomes (1KG) Project (phase 3) (7) and subsets of SNVs curated from The Exome Aggregation Consortium. The disease-related data set includes SNVs from The Human Gene Mutation Database (HGMD) (5) and pan-cancer data set (46) comprising publicly available somatic SNVs from The Cancer Genome Atlas (TCGA) (47), The Catalogue of Somatic Mutations in Cancer (48) and the SNV data set available from Alexandrov *et. al* (49). 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 data set. Furthermore, we removed HGMD variants in the 1KG and ExAC data sets. Similarly, we also removed known TCGA variants in the original ExAC SNV data sets.

SNVs from the pan-cancer data set were further sub-classified based on whether they fall in driver or passenger genes. Driver genes were curated from Vogelstein *et. al.* (50), in which driver and passenger genes are distinguished on the basis of mutational patterns. They define a driver gene as an oncogene if the SNV is recurrent at the same gene loci, whereas TSGs are mutated throughout their length. Similarly, we sub-classified passenger genes into cancer-associated genes (CAGs) and non-cancer associated genes (non-CAGs). CAGs include genes from the cancer gene census (CGC) (51) and a curated list of 4050 genes from a previous study (52). Furthermore, we removed any driver genes present in the CAG data set. The remaining set of genes impacted by pan-cancer SNVs constitutes our non-CAG data set.

Semi-balanced SNV data sets

The limited and uneven structural coverage of the human proteome primarily introduces two sources of potential bias when combined with SNV data sets: (i) some proteins may be over-represented when evaluating the effects of SNVs, and (ii) the sets of proteins that correspond to benign SNVs may differ considerably from those that correspond to deleterious SNVs, thereby complicating direct comparisons between benign and deleterious SNVs.

In order to address the first issue, we selected a non-redundant set of proteins within each data set. 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 1KG, ExAC and HGMD SNVs, respectively. Distributions delineating the number of SNVs within these non-redundant protein sets are given in Supplementary Figures S1–S3.

In order to address the second issue, we analyzed only those structures that fall within the *intersection* of the different non-redundant data sets. Thus, for each SNV in a structure within this intersection set of non-redundant proteins (which we term the 'semi-balanced set'), at least one residue overlaps with an ExAC/1KG and HGMD SNV. We utilized this semi-balanced set to demonstrate the utility of localized frustration (for evaluating deleteriousness) in the context of other methods (PolyPhen-2 (53) and SIFT (15)), as described in Results. We also compare ΔF distributions for 1KG, ExAC and HGMD SNVs on the semi-balanced set (Supplementary Figure S4).

Workflow for calculating ΔF

As mentioned, we investigated the impacts of different categories of SNVs on the local stability of protein structures. Quantifying the ΔF value associated with a given SNV at position i within a structure involves three steps (see also detailed formalism below): (i) determining F_{nat}: this represents a normalized energetic difference between having the wild-type residue at position i (the wild-type residue is generically termed 'nat' here) and the mean energy of having all 20 amino acids at position i, where all 20 of these energies are calculated using the wild-type structure; (ii) determining F'_{mut}: this represents a normalized energetic difference between introducing the *mutated* residue at position i (the residue resulting from the SNV is generically termed 'mut' here) and the mean energy of having all 20 amino acids at position i, where all 20 of these energies are calculated using a model of the mutated structure, which is built using the SNV; and (iii) determining the ΔF value associated with this SNV – this is simply the difference between the values calculated in steps (i) and (ii) above: $\Delta F = F'_{mut} - F_{nat}$. Qualitatively, ΔF thus describes the extent to which an amino acid change perturbs the local energetic landscape. We note that, in producing the model of the mutated structure in part (ii), this workflow performs a second-order calculation. As such, it provides values that may more accurately reflect the influence of introducing an SNV.

To map SNVs onto protein structures, the Variant Annotation Tool (VAT) (54) 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 identities. We then integrated the VAT annotation with the Biomart-derived (55) human gene and transcript IDs to map the SNV to specific structures in the protein data-bank (PDB). We restricted this SNV mapping scheme to high-quality structures with resolution values that were better than 2.8 Angstroms. Following SNV mapping to PDB structures, we generated models of the resultant mutated structures by applying homology modeling using the mutated protein sequence and the wild-type protein structure as input to Modeller (56,57).

Finally, we quantify the frustration index of the mapped residue in the wild-type structure as well as in the mutated model of the protein. Briefly, the residue-level localized frustration index (45) quantifies the degree to which that amino
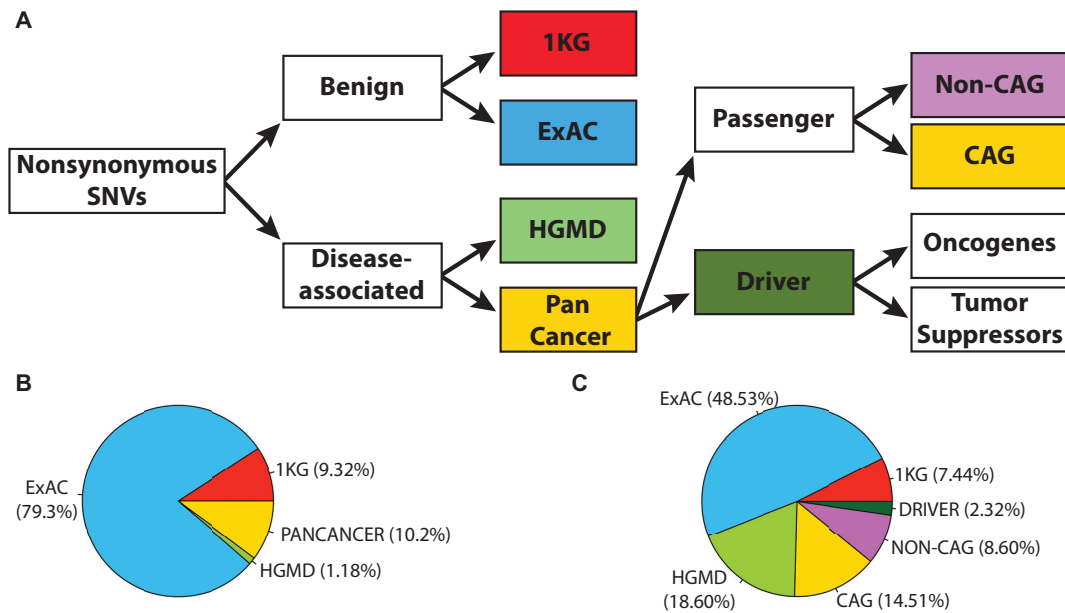
Figure 1. Overview of SNV categories and their relative proportions within the analyzed data pool. (**A**) Flowchart representing the different categories and origins of the variants analyzed in this study. A given non-synonymous SNV can be classified as benign or disease-related on the basis of its provenance (i.e. whether it is taken from 1000 Genomes, ExAC, HGMD or Pan-cancer variant data sets). Relative proportions of SNVs from various data sets (**B**) *prior* to and (**C**) *after* mapping SNVs to high-resolution protein databank (PDB) structures.

acid favorably contributes to the energy of the system relative to all 20 possible amino acids at that position:

$$F_i = \frac{\langle E_i^{T,\,U}\rangle - E_i^{T,\,N}}{\sqrt{1/N \sum_{k=1}^{n}(E_i^{T,\,U} - \langle E_i^{T,\,U}\rangle)}}$$, where $E_i^{T,\,N}$ is the total

energy of the wild-type protein. This energy is calculated using a function that includes an explicit water interaction term, $E_i^{T,\,N} = \sum_{k \neq i}^{n}(E_{\text{contact}}^{i;k} + E_{\text{water}}^{i;k}) + E_{\text{burial}}^{i}$. This water-mediated potential (44), describes the energies associated with direct interactions between residues i and k ($E_{\text{contact}}^{i;k}$), as well as those with water-mediated interactions between residues i and k ($E_{\text{water}}^{i;k}$) and an energy term associated with the burial of the residue (E_{burial}^{i}). The average energy of the decoy conformations ($\langle E_i^{T,\,U}\rangle$) is generated by mutating the original residue i to each of the alternative possible 19 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 values for large numbers of variants. Our benchmark calculations on 10,000 non-synonymous SNVs indicate that we can map, build mutated models and calculate ΔF values in ∼29 h on a single-core processor; specifically, we used an E5-2660 v3 (2.60GHz) processor. This approach is substantially more computationally tractable relative to traditional molecular dynamics simulations. Thus, it can readily be used to evaluate the effects of large numbers of SNVs. We also provide source code for the workflow on our GitHub page (https://github.com/gersteinlab/frustration).

In Figure 2, we demonstrate an example case in which a tryptophan residue at locus 31 within plastocyanin (PDB ID 3CVD) is mutated to tyrosine. For the wild-type 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, at this stage, the structure is not altered or minimized in any way). In this case, the energy computed using the wild-type residue (E_{TRP}) is substantially lower than the mean value $\langle E\rangle$ (rendering a positive value for ΔE_{TRP}). Because ΔE_{TRP} is greater than 0, the wild-type residue is said to be 'minimally frustrated'.

This same protein is known to contain a disease-related SNV at locus 31 (W31Y). To quantify the associated change in frustration, we first introduce tyrosine at locus 31 *in silico*, and then use Modeler to generate a model of the mutated structure. Thus, we now not only change the residue at locus 31, but also the configuration of the entire protein; the new structure is the model of the mutated protein. 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 ($\Delta E'_{\text{TYR}} < 0$), suggesting that the mutated residue is 'maximally frustrated'. We are primarily interested in the difference between these two states (ΔF). ΔF is proportional to the difference between $\Delta E'_{\text{TYR}}$ and ΔE_{TRP}. $\Delta \Delta E$ is defined to be the difference between the two energetic disparity measures ($\Delta \Delta E = \Delta E'_{\text{TYR}} - \Delta E_{\text{TRP}}$). Here, $\Delta \Delta E$ is less than 0, suggesting that the frustration value is higher in the mutated structure than that of the wild type.

Downstream analyses

In order to investigate the differential effects of SNVs in various data sets, we 'bin' each SNV into distinct categories based on their frustration indices and relative accessible surface areas (RSASA) in the wild-type structure. SNVs are classified into three groups (all in wild-type structures): (i)
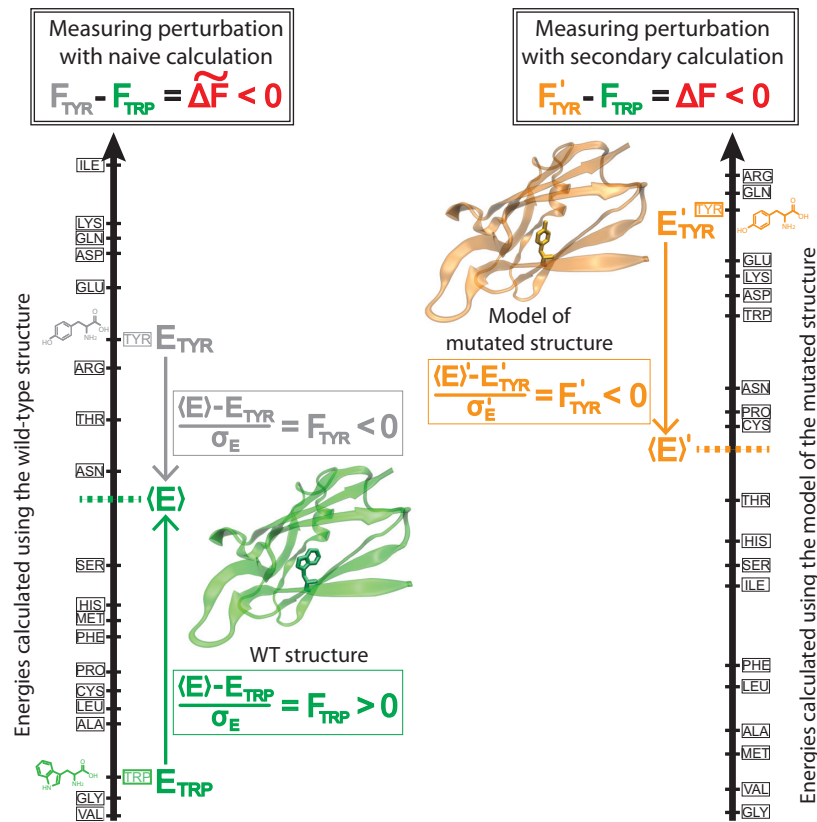
Figure 2. The effect of introducing a typical deleterious SNV ($\Delta F < 0$). Each of the two vertical lines represents an energy-level diagram. Each level on this energy scale corresponds to the total energetic value of the protein if the residue position (here, residue ID 31) were to be occupied by distinct amino acids (thus, for instance, in 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). 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 $(\langle E \rangle - E_{TRP})/\sigma_E = F_{TRP} > 0$. The gray vertical arrow designates the relative energy (E_{TYR}) associated with having a non-WT residue (TYR) at position 31 – this may conceivably be used to calculate F_{TRP} in a similar manner. Relating F_{TRP} to F_{TYR} by using the wild-type structure alone as a basis for estimating the perturbation induced by the W31Y SNV ($F_{TYR} - F_{TRP} = \tilde{\Delta}F$) would provide a naïve estimate of the perturbation. A more accurate evaluation of the perturbation requires a better estimate for F_{TYR}. This entails a secondary calculation, wherein the entire protein structure (right) is first changed (see Materials and Methods) by generating a model of the mutated structure after the SNV W31Y is introduced. This changed structure redistributes the relative energies 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 ($\langle E \rangle'$) of all possible 20 amino acids at that position. This disparity is designated by $(\langle E \rangle' - E'_{TYR})/\sigma_E' = F'_{TYR} < 0$. Taken together, the negative value associated with the disparity between the F'_{TYR} and F_{TRP} values ($F'_{TYR} - F_{TRP} = \Delta F < 0$) indicates that this SNV is locally unfavorable, and greater accuracy is likely obtained using the secondary calculation.

minimally frustrated (MinFNS); (ii) maximally frustrated (MaxFNS) and (iii) neutral (NeutFNS). MinFNS residues have frustration indices ≥ 0.78, whereas MaxFNS residues have frustration indices less than or equal to -1.0. SNVs falling between these two extremes fall 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 (58). Residues were defined to be in the core if their RSASA value $\leq 25\%$; surface residues had RSASA values $> 25\%$.

Furthermore, we investigated the differential influence of common and rare variants. SNVs with minor allele frequencies (MAF) $\leq 0.5\%$ were considered to be rare. SNVs were otherwise classified as common. Similarly, we also compared the effects of SNVs influencing the conserved and variable regions of the genome. The distinction between conserved and variable regions was defined using genome evolutionary rate profiling (GERP) scores (59). GERP scores are used to identify functionally constrained genomic elements based on multiple sequence alignments 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 scores ≤ 2.0 were considered to designate variable genomic loci. For each variant dataset, Table 1 provides a survey of the SNV counts across these different categories, along with the frequencies with which these variants lie in core or surface regions.

The deleteriousness of an SNV is a continuous variable, and indeed, this is reflected in the continuous nature of ΔF values. However, there is considerable value in applying a binary classification scheme to newly discovered SNVs, which

Table 1. Summary statistics on the number of SNVs used in comparative analyses

Conservation measure	1000 Genomes		ExAC	
	core	surface	core	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		SNV type	PANCAN	
	core	surface		core	surface
			non-CAG	2153	1848
GERP conserved (variable)	5158 (961)	1113 (221)	CAG	4140	2767
			driver	877	486

Shown are SNV counts for non-disease (top), HGMD (bottom-left) and pan-cancer SNVs (bottom-right). Variants were further classified as being rare (minor allele frequencies (MAF) $\leq$ 0.5%) or common (MAF > 0.5%), as well whether or not SNVs lie within conserved (GERP > 2.0) or variable (GERP $\leq$ 2.0) genomic regions.

may be predicted to be benign or deleterious. In order to perform such a binary classification, we applied a simplified decision boundary scheme, wherein we analyzed ΔF distributions for HGMD (disease-related) and ExAC (seemingly benign) variants. The threshold was set with the objectives of (i) minimizing the fraction of HGMD SNVs with ΔF values *above* the threshold, and (ii) minimizing the fraction of ExAC SNVs with ΔF values *below* the threshold. Using this approach, we observed that variants with ΔF scores $\leq$ −1.221 can be considered deleterious. Details of this scheme are provided as part of the Supplementary Material.

RESULTS

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

We performed a comparative analysis to investigate the impacts of benign (1KG and ExAC) and disease-related (HGMD) SNVs on the ΔF profiles of mutated residues in a large number of proteins. As detailed in Materials and Methods, each SNV data set was divided into three distinct categories based on the frustration index of the wild-type residue. Maximally frustrated residues in the wild-type 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 2; see also Materials and Methods for more details). For a given SNV, two protein structures are used in our analysis: the wild-type 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 wild-type and mutated model structures, respectively). Subsequently, we determine the difference between the frustration index of the native residue in the wild-type 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 in Materials and Methods). We observed that most SNVs (across all data sets) affecting maximally frustrated residues in the wild-type 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 base-substitution, thereby, leading to a negative frustration changes ($\Delta F < 0$) in a majority of cases across each data set. However, we emphasize that losses or gains in favorable interactions are dependent on the type of SNV (benign or disease-related) as well as whether the SNV affects 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 wild-type structure, and this trend is observed when using both ExAC and 1KG data sets (*P*-value < 2e-16 from two-sample Wilcoxon test; Figure 3A and B). In addition, disease-related 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 3C).
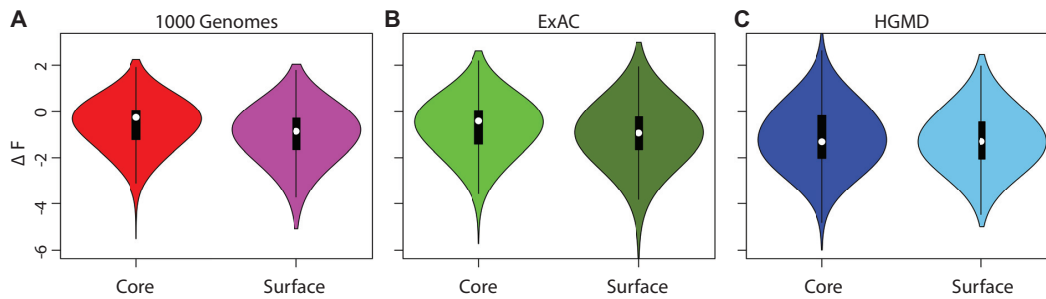
Figure 3. Differential effects of 'benign' and disease-associated SNVs on the localized frustration of minimally frustrated residues in the wild-type state. Violin plots showing ΔF distributions associated with SNVs affecting the core or surface, with SNVs taken from (**A**) 1000 Genomes, (**B**) ExAC and (**C**) HGMD. Comparisons between ΔF distributions for core and surface residues of the 1000 Genomes and ExAC data sets indicate that favorable interactions of surface residues in the wild-type states are highly disrupted upon mutation relative to core residues. Furthermore, the distributions of ΔF values for HGMD core residues were highly negative compared to 1KG and ExAC variants impacting core residues. The white dots, black boxes and vertical lines represent the medians, interquartile ranges and 95% confidence intervals of the ΔF distributions, respectively.

Differential effects of rare and common SNVs on localized frustration

In population-level studies, SNVs with lower MAF are generally interpreted as being more likely to be deleterious than SNVs with higher MAF values. Thus, within the set of benign SNVs (1KG and ExAC variants), MAF may be used as a proxy for varying degrees of selective constraint. Thus, we compared the ΔF distributions between rare and common SNVs. Consistent with our earlier observations regarding benign SNVs, we found larger disruptions to favorable local interactions in surface residues relative to core residues (Figure 4A). However, this disparity was slightly more pronounced for rare SNVs compared to common SNVs. This observation was consistent for the 1KG (Figure 4A) and ExAC data sets (Figure 4B; *P*-value < 2e-16 from two-sample Wilcoxon test). Furthermore, using both of these data sets, we observed that greater ΔF values associated with the introduction of SNVs (in either the positive or negative direction) tend to be associated with lower MAF values (Figure 4C, top and bottom panels). This trend is observed for SNVs that occur both on the surface and within the core.

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

We also examined the local perturbations induced by disease-related and benign SNVs in conserved and variable regions of the genome. We plotted distributions for the ΔF values for the surface and core residues (Figure 5), and observed that benign SNVs originating in both conserved and variable regions had similar effects on minimally frustrated core residues (Figure 5A and B). This observation was true for the surface residues as well. In contrast, disease-related SNVs intersecting conserved and variable genomic regions lead to variable ΔF values for surface residues (*P*-value = 4.715e-03 from two-sample Wilcoxon test). This disparity is even more pronounced in core residues (Figure 5C; *P*-value = 6.723e-09 from two-sample Wilcoxon test).

Differential effects of SNVs on driver and passenger genes

One of the most important challenge 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 (60). 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 (50), CAGs (52) and non-CAGs in the context of frustration. We mapped the somatic pan-cancer SNVs in these three distinct gene categories to protein structures, and then evaluated the ΔF distributions in all three.

As with benign SNVs, we observed that somatic SNVs in CAGs and non-CAGs generally lead to greater disruptions in minimally frustrated surface residues relative to core residues (Figure 6; *P*-value < 2.2e-16 from two-sample Wilcoxon test). This disparity was more pronounced among non-CAGs than CAGs (Figure 6). In contrast, SNVs in driver genes lead to larger disruptions in favorable localized interactions for surface and core residues (Figure 6; *P*-value < 2.2e-16 from two-sample Wilcoxon test) compared to CAG core and surface residues.

Differential effects of SNVs on oncogenes and TSGs

Cancer driver genes are classified as oncogenes and TSGs based on their mutational pattern and their mode of inducing tumorigenesis (50). Oncogenes are marked by recurrent SNVs within the same gene loci across different cancer types, and are believed to drive cancer progression through GOF mechanisms. In contrast, a TSG 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 7C, 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 7C, right panel).
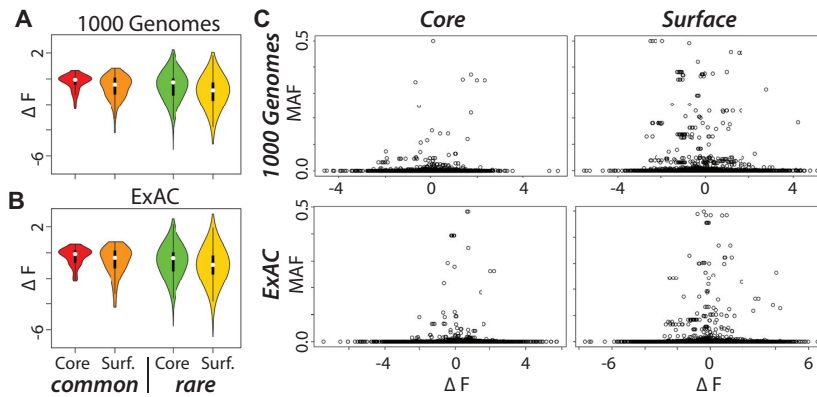
Figure 4. Common and rare SNVs differentially influence ΔF profiles of minimally frustrated core and surface residues. Violin plots showing the ΔF distributions induced by common and rare variants from in the (**A**) 1000 Genomes and (**B**) ExAC data sets (shown are the effects on minimally frustrated core and surface residues). Rare variants in both data sets 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 values (i.e. rarer SNVs). The white dots, black boxes and vertical lines represent the medians, interquartile ranges and 95% confidence intervals of the ΔF distributions, respectively.
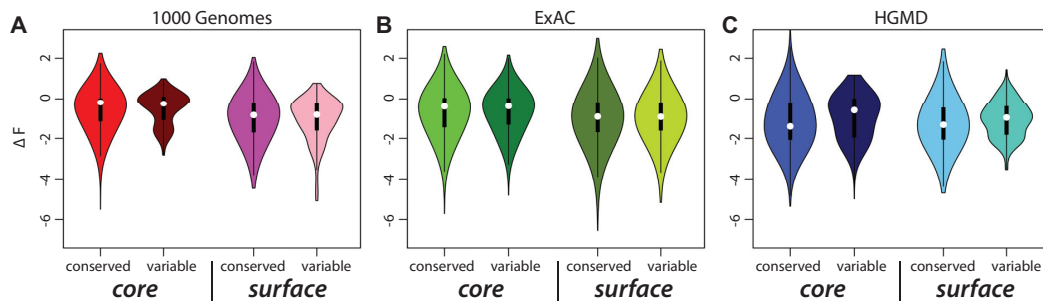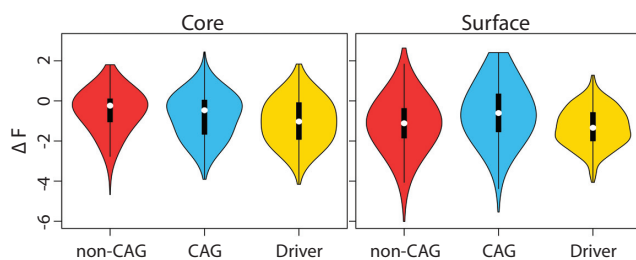

Figure 5. Comparisons between ΔF distributions for 'benign' and disease-related SNVs on conserved and variable residues. Violin plots depicting ΔF distributions introduced by (**A**) 1000 Genomes, (**B**) ExAC and (**C**) HGMD variants, respectively. ΔF distributions associated with HGMD SNVs indicate larger disruptions of conserved core residues compared to variable residues. In contrast, for the 1000 Genomes and ExAC data sets, no significant difference in ΔF distributions was observed for conserved and variable core residues (the same was true for surface residues). The white dots, black boxes and vertical lines represent the medians, interquartile ranges and 95% confidence intervals of the ΔF distributions, respectively.

Figure 6. 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) In contrast, SNVs in CAGs and non-CAGs disrupt favorable interactions of surface residues to a larger extent relative to core residues. The white dots, black boxes and vertical lines represent the medians, interquartile ranges and 95% confidence intervals of the ΔF distributions, respectively.

To evaluate the extent to which such effects manifest in our set of TSGs and oncogenes, we applied frustration to

evaluate changes in local perturbation when SNVs impact these distinct categories of driver genes (Figure 7A and B). We observed that SNVs in TSGs induce stronger perturbations in minimally frustrated core residues relative to surface residues (Figure 7A; *P*-value = 4.765e-03 from two-sample KS test). In contrast, SNVs in oncogenes induce greater ΔF values within minimally frustrated residues in the surface relative to core residues (Figure 7B; *P*-value = 1.91e-13 from two-sample KS test). 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 disruptions in favorable local interactions compared to oncogenes affecting driver SNVs in core residues (*P*-value = 6.753e-13 from two-sample KS test).

Localized frustration as a means of complementing global metrics

As discussed, existing structure-based methods for predicting SNV deleteriousness rely on global metrics of protein stability. These approaches may incorrectly predict known
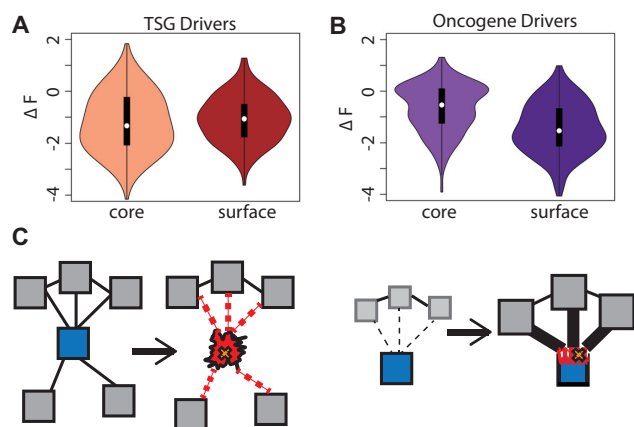
Figure 7. Differential impacts on ΔF distributions associated with SNVs on driver and passenger genes. Violin plots depicting ΔF distributions associated with (**A**) SNVs in TSGs and (**B**) oncogenes. Relative to surface residues, SNVs in TSGs lead to larger disruptions for minimally frustrated core 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 loss-of-function (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, black boxes and vertical lines represent the medians, interquartile ranges and 95% confidence intervals of the ΔF distributions, respectively.

disease-related 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 Materials and Methods), and predicted the impacts of these SNVs using SIFT, PolyPhen-2 and ΔF values. 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).

SIFT produces false negatives for 13.7% of these HGMD SNVs. We find that employing ΔF rescues 46% of these SIFT false negatives (i.e. by correctly predicting deleterious impacts). Similarly, PolyPhen-2 produces false negatives for 10% of the HGMD SNVs. Applying ΔF enables us to rescue 38% of these PolyPhen-2 false negatives. Glucokinase is used as an example to demonstrate specific cases of rescued variants (Supplementary Figure S6). A list of all false negatives rescued by ΔF analysis is provided in Supplementary Material data file. Finally, we also provide the full library of all HGMD SNVs mapped to protein structures (along with their corresponding PolyPhen-2, SIFT and ΔF scores), which may then be used as a resource in related studies.

DISCUSSION

Over the course of the last decade, tremendous improvements in sequencing and structural biology techniques have led to immense growth in genomic variation and three-dimensional structural data sets. This provides us with an ideal platform to investigate the impacts of genomic variants on protein structures. The objective of these studies is to gain mechanistic insights into the origin of various dis-

eases, as well as to design effective drug targets. Prior studies in this direction were limited by a lack of genomic variation and structural data. Moreover, these studies primarily focused on investigating the impacts of SNVs on the *global* stability of protein structure. However, many experimental studies have indicated a causal role for SNVs inducing local perturbation in various diseases. Here, 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. 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 X-ray structures in the PDB (11). Furthermore, cryo-EM is opening entirely new avenues for revealing the architectures of many proteins that were previously elusive to crystallography, which is expected to expand the structurally-resolved proteome (61). Finally, inferring how a given SNV affects a particular structure is not 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 (e.g. the consequences of an SNV within a multi-protein complex may adversely affect all members of the complex, despite the fact that the SNV maps to only one protein).

We employed an extensive catalogue of benign (∼5.7 million) and disease-related (∼0.76 million) SNVs. The benign SNV data set comprised SNVs from the 1KG Project (phase 3) and the ExAC Consortium. In contrast, HGMD SNVs and pan-cancer somatic SNVs constituted our disease-related SNV data set. We mapped ∼0.2 million benign and disease-related SNVs onto ∼10K high-resolution protein structures. Subsequently, we compared the impact of benign and disease-related SNVs on the frustration profiles of minimally frustrated residues in various protein structures. The ΔF distributions indicate that both benign and disease SNVs disrupt minimally frustrated surface residues to similar extents. However, the mechanistic differences between benign and disease SNVs can be attributed to their impacts on the local environment of core residues. Within the core, disease-related 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-related SNVs.

Furthermore, we quantified the influence of rare and common SNVs (present in healthy human populations) on the frustration profiles of affected protein residues. We observed that rare SNVs lead to larger local perturbations of minimally frustrated surface residues compared to common SNVs. This observation is intuitively consistent, as one would expect rare SNVs to have greater impacts on pro-

tein stability. In addition, we also investigated the differential impacts of SNVs intersecting conserved regions compared to variable genomic regions. The distinction between conserved and variable regions was based on GERP scores, which quantifies a cross-species conservation score on each nucleotide position. This cross-species conservation analysis indicated that there is no disparity between ∆F associated with benign SNVs 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 structures. In contrast, disease-related SNVs in conserved and variable regions of the genome exhibit significant differences in ∆F 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 relative to variable regions of the genome.

In addition to studying disease-related variants in general, tremendous progress in next-generation sequencing has led to unprecedented efforts to characterize cancer genomes. A great deal of effort has 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 in driver and passenger genes. Specifically, we studied these effects in the context of the local stability of protein structures. Our analysis indicates that SNVs influencing non-CAGs and CAGs lead to greater perturbations of surface residues compared to core residues. In contrast, SNVs that impact driver genes have similar affects on ∆F values in core and surface residues. These observations support 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 or TSGs. SNVs in oncogenes generally lead to cancer progression through GOF mechanisms, whereas SNVs impacting TSGs generally contribute to cancer growth through LOF events. These two distinct modes prompted us to closely inspect SNVs originating in each category, and we compared the ∆F profiles for SNVs across both. We observed that SNVs in oncogenes and TSGs generate greater ∆F values at the surface and core, respectively.

Comprehensive catalogues of SNVs from large-scale genomics projects have clearly established the important roles of disease-related and rare variants in human populations. We foresee further growth in genomic data sets as large-scale genomic consortia (such as the International Cancer Genomics Consortium, The Pan-Cancer Genome Atlas, The UK10K Project and the 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. 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 earlier studies that primarily employed global metrics (such as folding free energy changes) to quantify the effects of SNVs. We believe that the combination of these global and local metrics, along with sequence features, will further help to elucidate the mechanisms as well as predict the impacts of genomic variants.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

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. and Snyder,M.P. (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**, D514–D517.
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**, D980–D985.
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**, 64–69.

10. Lek,M., Karczewski,K.J., Minikel,E. V, Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

11. 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.

12. Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

13. 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–E464.

14. 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–D356.

15. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

16. Adzhubei,I.A. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

17. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **76**, 7.20.1–7.20.41.

18. 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.

19. 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**, 664–671.

20. 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.

21. 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*, **79**, 830–838.

22. 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.

23. 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 PPARγ F360L mutant associated with familial partial lipodystrophy. *Acta Crystallogr. D Biol. Crystallogr.*, **70**, 1965–1976.

24. 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.

25. 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**, e31677.

26. 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**, e77453.

27. 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.

28. Zhang,Z. (2013) A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. *Hum. Mol. Genet.*, **22**, 3789–3797.

29. Tsai,C.-J. and Nussinov,R. (2014) The free energy landscape in translational science: How can somatic mutations result in constitutive oncogenic activation? *PCCP*, **16**, 6332–6341.

30. 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.

31. 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*, **24**, 826–837.

32. 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**, 19819–19824.

33. 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**, W348–W351.

34. 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.

35. 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.

36. Clementi,C. and Plotkin,S.S. (2004) The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci.*, **13**, 1750–1766.

37. 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**, 171–180.

38. Frauenfelder,H., Sligar,S. and Wolynes,P. (1991) The energy landscapes and motions of proteins. *Science*, **254**, 1598–1603.

39. Camilloni,C. and Sutto,L. (2009) Lymphotactin: How a protein can adopt two folds. *J. Chem. Phys.*, **131**, 245105.

40. 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**, 3499–3503.

41. 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.

42. 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**, 12570–12575.

43. Changeux,J.-P. (2013) 50 Years of Allosteric Interactions: The twists and turns of the models. *Nat. Rev. Mol. Cell Biol.*, **14**, 819–829.

44. Zhuravlev,P.I. and Papoian,G.A. (2010) Protein functional landscapes, dynamics, allostery: A tortuous path towards a universal theoretical framework. *Q Rev. Biophys.*, **43**, 295–332.

45. Ferreiro,D.U., Komives,E.A. and Wolynes,P.G. (2013) Frustration in Biomolecules. *Q. Rev. Biophys.*, **47**, 1–97.

46. Davoli,T., Xu,A.W., Mengwasser,K.E., Sack,L.M., Yoon,J.C., Park,P.J. and Elledge,S.J. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**, 948–962.

47. 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**, 1113–1120.

48. 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.

49. 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**, 415–421.

50. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer Genome Landscapes. *Science*, **339**, 1546–1558.

51. 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.

52. 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.

53. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

54. 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.

55. 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–W598.

56. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

57. Webb,B. and Sali,A. (2014) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics*, **47**, 5.6.1–5.6.32.

58. 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.

59. 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.

60. 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.

61. 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.