Quantifying impact of non-synonymous coding variants on protein folding landscapes

Abstract

Next generation sequencing initiatives have lead to plethora of exome and genome sequence data. A major focus of these projects has been to identify and characterize the effect of genomic variations in coding region. Previous studies have quantified these effects of non-synonymous coding single nucleotide variants (SNVs) on the global stability of the protein structure. Numerous experimental studies implicated important roles of non-synonymous coding SNVs in local perturbation of protein structures, leading to various diseases. In this work, we describe a workflow to evaluate localized frustration index as a metric to investigate local changes in protein structure upon mutation. We applied our workflow on a large number of benign and disease non-synonymous SNVs. Subsequently, we evaluated differences between benign and disease mutations and found that localized frustration changes within proteins is closely related to the deleteriousness associated with SNVs. Furthermore, we observed that rare variants in healthy individuals tend to disrupt local interactions to a larger extent compared to common variants. Moreover, frustration characterization for mutations in oncogene and tumor suppressor genes (TSGs) indicated that oncogenes disrupt the surface residues to a larger extent, thereby facilitating gain-of-function events. In contrast, SNVs impacting TSGs locally disrupt core residues, and thereby potentially induce loss-of-function effects.

Deleted: However, numerous

Deleted: present

Deleted: novel

Deleted: disruptions of favorable local interactions play key roles in determining

Deleted: of mutations.

Deleted: benign

Deleted: perturb

Deleted: drive cancer progression

Introduction

The advent of next-generation sequencing technologies has led to a remarkable increase in genomic variation data at exome as well as whole-genome levels [1, 2]. These large genomic variation datasets are playing a pivotal role in advancing efforts toward personalized medicine[3]. Small non-synonymous coding SNVs are of particular interest due to their clinical relevance [4, 5, 6]. This has led toward a greater effort to curate disease-associated non-synonymous coding 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]. Moreover, initiatives such as The 1000 Genomes Project [7, 8], Exome Sequencing Project (ESP) [9] and Exome Aggregation Consortium (ExAC) [10] have generated catalogue of coding SNVs in normal human populations.

Formatted: Font color: Black

As the costs associated with sequencing entire human genomes and exomes continue to fall, sequencing will become routine in both medical and academic contexts \cite {26658741}. Indeed, it may take less than a decade to reach the milestone of a million sequenced genomes, resulting in massive datasets of rare SNVs. This exponential growth in the number of newly discovered rare SNVs pose significant challenges in terms of variant interpretation. 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. 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.

Simultaneously, immense progress has been made in resolving the three-dimensional structures of many proteins over the last several decades [11]. A large amount of protein-protein, protein-ligand and protein-nucleic acid complex structures have been solved at high resolution. This complementary evolution of sequence and structural space provides an ideal platform to investigate the functional and structural consequences of benign and disease SNVs on protein structures. Integration of these

growing knowledge-bases of SNVs and three-dimensional structures will lead to greater understanding of the biophysical mechanisms behind various diseases. In addition to gaining a better understanding of how disease-SNVs impart deleterious effects, these integrative frameworks can be utilized to both predict the impacts of poorly understood SNVs (for which disease association is unknown) and to prioritize SNVs based on predicted deleteriousness [12, 13, 14, 15]. We also note that this integration of structural information with sequence data to infer mechanistic effects will aid in more intelligent and targeted design of drugs in various therapeutic contexts.

The integration of genomic variants and structure data is itself not new: prior studies have assessed the impact of disease and benign SNVs by evaluating changes in protein stability, whereby the extent to which a mutation disrupts thermodynamic stability is analyzed [16, 17, 18]. These approaches rely on the fact that SNVs induce changes in the folding landscape and conformational ensemble of a protein. Such changes in global thermodynamic stability are often quantified by calculating the folding free energy change ($\Delta\Delta G$) after mutating the relevant residue in the protein structure [18, 19].

Importantly, however, many disease-associated SNVs introduce *local* structural changes without appreciably affecting the folding free energy [20, 21]. Such local perturbations may include disruptions of hydrogen bond networks [22, 23] and salt bridges [24, 25]. In most of these cases, overall stability of the protein is not drastically affected. Examples of such local yet impactful events may include changes in catalytic centers, "hotspot residues" within protein-protein interaction interfaces, or key allosteric residues. Changes in such residues may impart only minimal effects to the protein's overall architecture, but may nevertheless ablate proper functioning entirely. The important role of disrupted local interactions in many diseases mandates closer inspection of SNVs induced localized perturbation in protein structures. We examine the role of localized perturbations by calculating changes in the *localized frustration index* [26, 27] of residues impacted by non-synonymous SNVs.

The concept of *localized frustration* was originally introduced by Wolynes *et al.* to describe the protein folding landscape[26]. The protein folding process is believed to follow a smooth funneled energy landscape, where strong energetic conflicts are avoided [28, 29]. The absence of conflicting interactions is often considered as presence of

Formatted: Font color: Black

Deleted: global

Deleted: there are no significant changes in the global structure

Deleted: but

Deleted: perturbations in hydrogen-bond network

Deleted: salt bridges play an important role in driving the disease.

Deleted: progression of various

Formatted: Font color: Black

minimal frustration on the protein folding landscape [30, 31]. Furthermore, it becomes the basis for the widely acclaimed "principal of minimal frustration" for protein folding, whereby the final native state of the protein is "minimally frustrated" and occupies the global minima on the energy landscape [32]. The native interactions are stronger in the global minimum conformation compared to random non-native interactions in a molten globule structure.

However, this principal of minimizing frustration globally does not preclude the possibility of local frustration, which is often considered essential to protein biology and function [33, 34, 35]. Ferriero *et al.* proposed a framework to compute the *localized frustration* profile of a given protein [26]. The *localized frustration index* quantifies the contribution of each residue or residue pair in the total energy of native structure compared to their energetic contribution in a random non-native configuration [36]. A native residue (residue pair) in a protein structure is considered *minimally frustrated* when it contains sufficient extra stabilization energy in its native state. In contrast, sufficiently destabilizing residue (residue pair) in the protein structure is considered *as maximally frustrated*. In addition, a residue (residue pair) is considered neutral when its stability profile lies in between these extremes.

In this work, we evaluated changes in *localized frustration index* to assess the local perturbations introduced by "benign" and disease-associated SNVs on the folding landscape of various proteins. In our analysis, we consider SNVs as "benign" and disease-associated if they are present in healthy (The 1000 Genome and EXAC projects) and ailing (HGMD) populations, respectively. We also applied this scheme on cancer somatic variants to differentiate the influence of mutations impacting driver and passenger genes. Majority of our analyses were consistent with prior studies exploring impact of SNVs on protein structure. However, we provide a distinct rationale behind these observations based on localized frustration framework. For instance, we observe that large disruption in local interactions of *minimally frustrated* core residues distinguishes "benign" from disease mutations as well as mutations impacting passenger and driver genes in cancer. In contrast, "benign" SNVs and passenger gene intersecting mutations generate larger perturbation in local interactions of *minimally frustrated* surface residues compared to core residues. Furthermore, comparison between rare and

Deleted: global stabilization

Deleted: its

Deleted: its molten globule configurations

Deleted: We

Deleted:

Deleted: we also compare the impact of

common non-synonymous SNVs within healthy human population indicated that rare variants induce larger disruption in favorable local interactions compared to common variants. Moreover, we investigated the effects of non-synonymous SNVs impacting conserved and variable regions of proteins, where conservation was measured across different species. For disease SNVs, we detected a significant disparity between local perturbations observed due to mutations impacting conserved regions compared to variable regions of proteins. However, no such disparity was observed for "benign" mutations. In addition, we also present new insights gained through the frustration framework while honing into the biological affects of cancer-associated SNVs. We compared changes in *local frustration* indices that are prompted upon the introduction of SNVs in oncogenes vs. tumor suppressor genes (TSGs). A closer inspection indicated that oncogenes and TSGs employ two distinct mode for cancer progression. Somatic mutations in oncogenes disrupt local interactions of *minimally frustrated* surface and facilitate cancer progression through large number of non-specific interactions. However, mutations in TSGs drive cancer progression through larger local perturbations in *minimally frustrated* core residues.

Results

Datasets of non-synonymous SNVs & their structural coverage

In order to evaluate the impact of various types of non-synonymous variants on frustration profile of protein residues, we collected and analyzed data from a variety of sources. These sources were chosen in order to obtain both benign and disease-associated SNVs, with the disease-associated SNVs having been further sub-classified to investigate disease mechanisms in greater detail. An overview of this data collection scheme is provided in Figure 1a. Figure 1b gives summary statistics on all non-synonymous SNVs in these datasets, and Figure 1c provides the corresponding data on the subset of these SNVs which were mapped to high-resolution protein structures from the PDB. Further details on the statistics obtained as part of this data collection framework are provided in *supplementary information*.

Deleted: populations.

Deleted: originating in

Deleted: regions of the genome,

Deleted: compare these to SNVs in

Deleted: originating in

Deleted: the genome. No

Deleted: among

Deleted: Further

Deleted: ,

Deleted: w

Deleted: mechanisms

Deleted: . ... [1]

Formatted: Line spacing: single

Formatted: Underline

Deleted: here

Deleted: We collected and annotated 6.46 million non-synonymous SNVs using VAT [49]. About 5.1 million of these SNVs were benign mutations that were obtained from the ExAC Project, and an additional roughly 0.6 million SNVs were taken from phase 3 of the 1000 Genomes Project, which constitutes 79% and 9% of our total set of annotated SNVs, respectively (Figure 1b). The remaining SNVs were a set of disease-associated mutations, and these comprised ~76,000 HGMD SNVs and 0.65 million publicly available pan-cancer somatic SNVs. HGMD and the pan-cancer dataset constituted 2% and 10% of the total collected non-synonymous SNVs, respectively (Figure 1b). ... [2]

Formatted: Indent: First line: 0"

Formatted: Font:14 pt, Bold, Font color: Black

Differential effects of benign and disease-associated SNVs on localized frustration

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

For each SNV, the change in frustration ($\Delta\Delta F$) was calculated as follows (see Figure 2). For a given SNV mapped to a PDB structure, then two protein structures are used in our analysis: the *native structure* (as it exists in the crystallized protein in the PDB), and the *non-native structure* (which 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 models, the *frustration index* is calculated at residue *j* (ΔF_{nat} and $\Delta F_{non\text{-}nat}$ for the native and non-native structures, respectively). Subsequently, we determine the difference between the *localized frustration index* of the wild-type residue in the native structure and the mutated residue in the non-native structure ($\Delta F_{non\text{-}nat} - \Delta F_{nat} = \Delta\Delta F$).

After calculating the difference between the *frustration index* ($\Delta\Delta F$) of a residue in the mutated model and the native structure in all three categories, the resultant distributions are plotted (further details are given under Methods). We observed that most SNVs (across all datasets) influencing *maximally frustrated residues* in the native structure induce small $\Delta\Delta F$ change. This suggests that changes *to maximally frustrated residue* alleviate the extent of conflicting interactions, thereby resulting in a positive frustration difference ($\Delta\Delta F > 0$). In contrast (and as expected), residues that are originally *minimally frustrated* lose favorable interactions leading to a negative frustration difference ($\Delta\Delta F < 0$) upon mutation in majority of cases across each dataset. However,

Deleted: If

Deleted: can be

Deleted: but positive

Deleted: become more frustrated

we emphasize that losses or gains in favorable interactions vary based on the type of mutation (benign or disease) as well as whether or not the mutation affects a residue on the surface or within the protein core.

We observed that "benign" SNVs lead to greater disruptions in favorable interactions for *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-values = 2e-16 and 2e-16 from two-sample Wilcoxon and KS tests, respectively*) (Figure 3a & 3b). In addition, disease-associated SNVs (from HGMD) result in similar *frustration changes* between core and surface residues (*p-values = 2e-16 and 2e-16 from two-sample Wilcoxon and KS tests, respectively*) (Figure 3c). However, SNVs from HGMD that impact *minimally frustrated* core residues disrupt the local interactions to a greater extent than "benign" mutations.

Differential effects of rare and common SNVs on localized frustration

In population level study, SNVs with lower derived allele frequency (DAF) are generally interpreted to be more deleterious than SNVs with higher DAF 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 abated ΔΔF distribution for *minimally frustrated* core and surface residues. Consistent with our earlier observation related to benign mutations, we found larger disruption in favorable local interactions for surface residues compared to core residues (see Figure 4a1 & 4a2). However, this disparity was slightly more pronounced for rare SNVs compared to common SNVs. This observation was consistent for the 1000 Genome (Figure 4a1) and EXAC datasets (Figure 4a2). Furthermore, using both of these datasets, we observed that greater changes in *frustration* associated with the introduction of SNVs (in either the positive or negative directions) tend to be associated with lower MAF values (Figures 4b1,4b2,4c1and 4c2). This trend is observed for SNVs that occur on both the surface and within the core.

Deleted: interestingly

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

We also examined the local perturbation induced by disease-associated and benign SNVs originating in conserved and variable regions of the genome. We plotted ΔΔF value distribution to perform this analysis for the surface and core residues, independently. Consistent with our earlier analysis, we found that benign surface mutations disrupt favorable local interactions to a greater extent compared to benign core mutations (Figure 5a). Furthermore, we observed that benign mutations originating in both conserved and variable regions of the genome had similar effects on *minimally frustrated* core and surface residues (Figure 5a & 5b). In contrast to benign SNVs, disease-associated SNVs intersecting with conserved and variable genomic regions lead to variable changes in frustration profile for surface residues (*p-values = 0.00031 and 8.146e-05 from two-sample Wilcoxon and KS tests, respectively*). This disparity is more pronounced in core residues (*p-values = 3.298e-08 and 2e-16 from two-sample Wilcoxon and KS tests, respectively*) (Figure 5c).

Differential effects of SNVs on driver and passenger genes

One of the most important challenges confronting the cancer genomics community involves discriminating between highly penetrative driver mutations from the large number of passenger mutations that arise over the course of tumor progression [50]. 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, cancer-associated genes (CAGs) and non-cancer associated genes (non-CAGs) in the context of *frustration* [42]. We mapped the somatic pan-cancer SNVs that intersect these these three distinct gene categories onto high-resolution protein structures. We then evaluated the ΔΔF distributions in all three categories.

As with benign mutations, we observed that somatic mutations impacting CAGs and non-CAGs lead to greater disruptions *in minimally frustrated surface* residues relative to *minimally frustrated core* residues (*p-values = 2.2e-16 and 2.2e-16 from two-sample Wilcoxon and KS tests, respectively*) (see Figure 6a). Moreover, this variability in ΔΔF distribution between core and surface residues was more pronounced among non-CAGs compared to CAGs (see Figure 6a & 6b). In contrast, SNVs that impact *driver genes* lead to larger disruptions in favorable localized interactions for surface (*p-values* =

Deleted: ,

Deleted: (driver, cancer-associated, and non-cancer associated genes are believed to be directly actionable, indirectly actionable, and unrelated to tumorigenesis, respectively – see [42]).

Deleted: indirectly actionable

Deleted: actionable genes

Deleted: actionable genes

Deleted: indirect actionable genes

Deleted: to what we observe in *indirectly actionable* and non-actionable genes

0.005 and 0.016 from two-sample Wilcoxon and KS tests, respectively) and core residues (*p-values = 2.2e-16 and 2.2e-16 from two-sample Wilcoxon and KS tests, respectively*) (see Figure 6c).

Deleted: . ... [3]

Formatted: Indent: First line: 0"

Differential effects of SNVs on oncogenes and tumor-suppressor genes

Cancer driver genes are classified as oncogenes and tumor suppressor genes based on the mechanisms by which they are believed to influence tumorigenesis [41]. Oncogenes are marked by recurrent mutations 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 non-synonymous SNVs that are scattered throughout the gene, and they are believed to facilitate cancer progression through loss-of-function (LOF) mechanisms. Thus, one may intuitively expect that cancer somatic variants within tumor-suppressor genes tend to be enriched within the buried core, whereas the gain-of-function effects induced in oncogenes are enriched on the surface. This line of thinking is guided by the idea that loss-of-function variants often act by destabilizing the entire protein (Figure 7a), whereas gain-of-function 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 7b).

In order to evaluate the extent to which such effects manifest in our set of tumor-suppressor genes and oncogenes, we measured the enrichment of non-synonymous SNVs in buried protein regions and compared these enrichment values to the number of SNVs expected to be present in buried regions under a null model, where SNVs are randomly (uniformly) distributed throughout the protein. In agreement with the model proposed above, we find that SNVs are enriched within buried regions of tumor-suppressors relative to this null (Figure 7c), whereas they are depleted in buried regions in oncogenes (Fig 7d). Finally, we note that this observation remains consistent even if one alternatively counts the affected *loci* within proteins (as oppose to the number of *SNVs*) (Figures [[non_redund_model_tsg.jpg]] and [[non_redund_model_oncos.jpg]] for tumor-suppressors and oncogenes, respectively.

Furthermore, we applied the *localized frustration* framework to evaluate changes in local perturbation when non-synonymous coding mutations impact these distinct categories of driver genes (Figure 8). We observed that SNVs affecting TSGs induce stronger perturbations in *minimally frustrated* core residues relative to surface residues (*p-values = 8.15e-2 and 4.7e-3 from two-sample Wilcoxon and KS tests, respectively*) (Figure 8a). In contrast, SNV affecting oncogenes induces greater changes in *localized frustration* within *minimally frustrated* residues in the *surface* relative to *minimally frustrated* residues in the *core* (*p-values = 2.2e-16 and 1.91e-13 from two-sample Wilcoxon and KS tests, respectively*) (Figure 8b). Moreover, SNVs impacting *oncogenes* lead to larger disruptions in favorable local interactions compared to *TSGs* for *minimally frustrated* surface residues (*p-values = 5.0e-4 and 2.3e-3 from two-sample Wilcoxon and KS tests, respectively*). However, mutations impacting *TSGs* lead to greater disruption in favorable local interactions compared to oncogenes affecting driver SNVs in core residues. (*p-values = 6.306e-15 and 6.753e-13 from two-sample Wilcoxon and KS tests, respectively*).

Formatted: Normal

Deleted: These two distinct modes of cancer progression motivated us to apply our framework to characterize

Deleted: 7). ... [4]

Formatted: Font:Italic

Deleted: 7b). Furthermore

Deleted: *TSGs*

Deleted: *oncogenes*

Deleted:) (Figure 7a). In contrast, SNVs affecting TSGs induce stronger perturbations in *minimally frustrated* core residues relative to surface residues (*p-values = 8.15e-2 and 4.7e-3 from two-sample Wilcoxon and KS tests, respectively*). Moreover, TSG influencing driver mutations

Formatted: Font color: Black

Discussion

In the last decade, tremendous improvements in sequencing and structural biology techniques have lead 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 are 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 information data. Moreover, these studies primarily focused on investigating the impact of non-synonymous SNVs on the *global* stability of protein structure. However, many experimental studies have clearly indicated causal role of coding SNV induced local perturbation in various diseases. In this work, we repurpose the concept of *localized frustration*, originally introduced in protein folding studies to quantify coding SNV-induced local perturbations. The *localized 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.

In this study, we employed an extensive catalogue of "benign" (~5.7 million) and disease-associated (~0.76 million) non-synonymous 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 frustration change (ΔΔF) distributions indicated that both "benign" and disease SNVs disrupt *minimally frustrated* surface residues to similar extents. However, the mechanistic difference between "benign" and disease mutations 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" mutations. 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 non-synonymous 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 institutively consistent as one would expect rare SNVs to have grater 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 ΔΔF associated with *benign* mutations originating in conserved and variable regions. This lack of disparity can be attributed to the absence of significant local perturbations induced by benign mutations, 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

Formatted: Indent: First line: 0.5"

Deleted: . Furthermore

observe significant differences in ΔΔF values. This is consistent with prior studies, which indicate that the deleteriousness of a mutation is more pronounced when non-synonymous 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 genome. Large efforts have been invested to discriminate between driver mutations from passenger mutations. Driver mutations 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 (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 induce comparable changes in frustration in core and surface residues. These observations further reiterate our earlier conclusion that the deleteriousness of a given mutation 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. Mutations in oncogenes lead to cancer progression through gain-of-function mechanisms, whereas SNVs impacting tumor suppressor genes contribute to cancer growth through loss-of-function mechanisms. These two distinct mode prompted us to closely inspect SNVs originating in oncogenes and TSGs. Our enrichment analysis indicated that SNVs impacting tumor-suppressors gene influence buried residues on protein structure whereas oncogenes SNVs predominantly affect surface residues. Furthermore, we compared the changes in *localized interaction profile* for residues influenced by these two distinct categories of SNVs. we observed that mutations in oncogenes and TSGs generate greater changes in *localized frustration* (ΔΔF) profile for *minimally frustrated* surface and core, respectively.

Formatted: Font:14 pt, Bold, Underline

Deleted: mechanisms

Deleted: We

Deleted: We

Deleted: This suggests that SNVs in TSGs destabilize the hydrophobic core by disrupting locally favorable interactions. In contrast, mutations in oncogenes lead to higher local perturbations in surface residues. These perturbations potentially drive carcinogenesis by facilitating non-specific protein-protein interactions.

Comprehensive catalogues of genomic variations from large-scale genomics project have clearly established the important role of disease-associated and rare variants in human populations. We foresee further growth in genomic variation datasets as large-scale genomic consortium projects such as International Cancer Genomics Consortium (ICGC), The Pan-Cancer Genome Atlas (PANCAN Atlas), UK10K project and Mendelian genomic program will continue to decipher mutational landscape of human genomes and exomes. Similarly, advancement in electron microscopy, NMR, single electron scattering and other biophysical 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 insight in disease progression and design effective drugs for disease treatment. In this work, we demonstrate the role of local perturbation 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 affects of genomic variants. We strongly believe that combination of these global and local metrics, along with sequence features, will help us elucidate the mechanism as well as predict the impact of genomic variations in disease and healthy human populations.

Deleted: 0

Method

Non-Synonymous SNV Datasets

We utilized a comprehensive catalogue of non-synonymous SNVs from various resources. Our non-synonymous dataset is divided into two broad categories (benign and disease-associated) (Figure 1a). The "benign" set comprises of SNVs reported in The 1000 Genome Project (phase 3) [7] and The Exome Aggregation Consortium [10]. Disease-associated dataset included SNVs from the Human Genome Mutational Database (HGMD) [5] and pan-cancer dataset [37]comprising of publicly available coding somatic mutations from The Cancer Genome Atlas (TCGA) [38], The Catalogue of Somatic Mutations in Cancer (COSMIC) [39] and coding SNV dataset available from Alexandrov

et. al [40]. 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. [41], where they distinguish between driver and passenger genes based on mutation pattern. They define a driver gene as an oncogene if the mutation 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) [42]and a curated list of 4050 genes from a previous study [43]. Furthermore, we removed any driver gene present in the CAG dataset. The remaining set of genes impacted by pan-cancer mutations constituted our non-CAG dataset.

Deleted: into

Deleted: Furthermore, they

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 changes in *localized frustration index* ($\Delta\Delta F$) of mutated residues to quantify SNVs induced local perturbation. Quantifying changes in localized frustration 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 frustration difference ($\Delta\Delta F$) of mutated residue in the native and mutated conformations.

To map SNVs onto protein structures, Variant Annotation Tool (VAT) \cite {22743228} was applied to annotate our curated catalogue of non-synonymous 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[44] derived human gene and transcript IDs to map the SNV on to specific PDB structures. We restricted this SNV mapping scheme to high-quality crystal 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 modeller [45, 46].

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 [36] quantifies the favorable stability imparted to a residue in the native structure compared to other possible residues at that particular location: $F_i = \frac{<E_j^{T,U}> - E_i^{T,N}}{\sqrt{1/N \sum_{k=1}^{n} (E_j^{T,U} - <E_j^{T,U}>)}}$, where $E^{T,N}$ 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, $E^{T,N} = \sum_{k \neq i}^{n} (E_{contact}^{i;k} + E_{water}^{i;k}) + E_{burial}^{i}$. This function, termed the associated water-mediated (AWM) potential [36], 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 ($< E_j^{T,U} >$) is generated by mutating 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.

In Figure 2, we demonstrate an example case in which replacing isoleucine at a particular locus within ubiquitin (PDB ID 1UBQ) with a tyrosine. Shown on the left (in green) is the native (i.e., wild-type structure). The vertical axis designates the different energies that would result when the residue at this locus is mutated each one of the other 19 amino acids. Specifically, these 19 decoy energies are only 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 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. In this case, the energy computed using the wild-type residue (ILE) is substantially lower than this mean value (rendering ΔE_{nat} greater than 0). Because ΔE_{nat} is greater than 0, this wild-type isoleucine is said to be *"minimally frustrated"*.

This same protein is known to contain a non-synonymous disease-associated SNV at locus X. Specifically, the disease-associated change occurs when the isoleucine is

mutated to tyrosine. To quantify changes in localized frustration of ILE in this example, we first introduce the tyrosine at locus Y *in silico*, and then use Modeller 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 Y, but also the configuration of the entire protein; the structure is that 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 Y is higher than the mean energy among all 20 amino acids within the non-native structure ($\Delta E_{non\text{-}nat} < 0$), suggesting that the mutated residue is *"maximally frustrated*". We are primarily interested in the frustration change between these two states. This value is proportional to the difference between $\Delta E_{non\text{-}nat}$ and ΔE_{nat}. ($\Delta E_{non\text{-}nat} - \Delta E_{nat} = \Delta\Delta E$) Here, $\Delta\Delta E$ is less than 0, suggesting that the frustration is higher in the mutated structure than that of the wild type.

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 (Table 1). SNVs are classified in three groups based on the frustration index of the mapped residue in 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 to be NeutFNS. 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 [47]. Residues were defined as core when the RSASA value was lower than or equal to 25 % and surface residues had RSASA value greater than 25%.

Furthermore, we investigated the differential influence of common and rare mutations in healthy human populations. Benign SNVs with derived allele frequency (DAF) 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 region of the genome. Distinction between

conserved and variable region of the genome were derived based on GERP score [48]. GERP score identifies functionally constrained elements in genome based on multiple sequence alignment of genomic sequences from multiple species. In our analysis, we defined a genomic position as conserved if its GERP score > 2.0. Similarly, we considered genomic location to be variable, when the GERP score was positive and less than or equal to 2.0(<= 2.0).

References

1. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011) The real cost of sequencing: higher than you think!. *Genome Biol* 12, 125, PMCID: 21867570.
2. Soon, W. W., Hariharan, M., & Snyder, M. P. (2013) High-throughput sequencing for biology and medicine. *Mol Syst Biol* 9, 640, PMCID: 23340846.
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., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A. P., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M. A., Greenberg, P. L., Snyder, P., Klein, T. E., Altman, R. B., Butte, A. J., Ashley, E. A., Gerstein, M., Nadeau, K. C., Tang, H., & Snyder, M. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293-307, PMCID: 22424236.
4. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33, D514-7, PMCID: 15608251.
5. Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized

Deleted: ... [5]

genomic medicine. *Hum Genet* 133, 1-9, PMCID: 24077912.
6. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42, D980-5, PMCID: 24234437.
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., & Abecasis, G. R. (2015) A global reference for human genetic variation. *Nature* 526, 68-74, PMCID: 26432245.
8. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, PMCID: 23128226.
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., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad GO, Seattle GO, & NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64-9, PMCID: 22604720.
10. Consortium, E. A. (2015) Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* , , PMCID: .
11. 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., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E., & Burley, S. K. (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 43, D345-56, PMCID: 25428375.
12. Ng, P. C. & Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31, 3812-4, PMCID: 12824425.
13. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248-9, PMCID: 20354512.

14. Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7.20, PMCID: 23315928.
15. Wong, W. C., Kim, D., Carter, H., Diekhans, M., Ryan, M. C., & Karchin, R. (2011) CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 27, 2147-8, PMCID: 21685053.
16. Zhang, Z., Wang, L., Gao, Y., Zhang, J., Zhenirovskyy, M., & Alexov, E. (2012) Predicting folding free energy changes upon single point mutations. *Bioinformatics* 28, 664-71, PMCID: 22238268.
17. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., & Alexov, E. (2013) Molecular mechanisms of disease-causing missense mutations. *J Mol Biol* 425, 3919-36, PMCID: 23871686.
18. Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79, 830-8, PMCID: 21287615.
19. Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A., & Böckmann, R. A. (2009) Predicting free energy changes using structural ensembles. *Nat Methods* 6, 3-4, PMCID: 19116609.
20. Lori, C., Pasquo, A., Montanari, R., Capelli, D., Consalvi, V., Chiaraluce, R., Cervoni, L., Loiodice, F., Laghezza, A., Aschi, M., Giorgi, A., & Pochetti, G. (2014) Structural basis of the transactivation deficiency of the human PPARγ F360L mutant associated with familial partial lipodystrophy. *Acta Crystallogr D Biol Crystallogr* 70, 1965-76, PMCID: 25004973.
21. Monticone, S., Bandulik, S., Stindl, J., Zilbermint, M., Dedov, I., Mulatero, P., Allgaeuer, M., Lee, C.-C. R., Stratakis, C. A., Williams, T. A., & Tiulpakov, A. (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, E114-8, PMCID: 25322277.
22. Doss, C. G. P. & Nagasundaram, N. (2012) Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. *PLoS One* 7, e31677, PMCID: 22384055.

23. Kumar, A., Rajendran, V., Sethumadhavan, R., & Purohit, R. (2013) Molecular dynamic simulation reveals damaging impact of RAC1 F28L mutation in the switch I region. *PLoS One* 8, e77453, PMCID: 24146998.
24. Boccuto, L., Aoki, K., Flanagan-Steet, H., Chen, C.-F., Fan, X., Bartel, F., Petukh, M., Pittman, A., Saul, R., Chaubey, A., Alexov, E., Tiemeyer, M., Steet, R., & Schwartz, C. E. (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, 418-33, PMCID: 24026681.
25. Zhang, Z., Norris, J., Kalscheuer, V., Wood, T., Wang, L., Schwartz, C., Alexov, E., & Van Esch, H. (2013) A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. *Hum Mol Genet* 22, 3789-97, PMCID: 23696453.
26. Ferreiro, D. U., Hegler, J. A., Komives, E. A., & Wolynes, P. G. (2007) Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci U S A* 104, 19819-24, PMCID: 18077414.
27. Jenik, M., Parra, R. G., Radusky, L. G., Turjanski, A., Wolynes, P. G., & Ferreiro, D. U. (2012) Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res* 40, W348-51, PMCID: 22645321.
28. Onuchic, J. N., Luthey-Schulten, Z., & Wolynes, P. G. (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 48, 545-600, PMCID: 9348663.
29. Chavez, L. L., Onuchic, J. N., & Clementi, C. (2004) Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc* 126, 8426-32, PMCID: 15237999.
30. Clementi, C., Nymeyer, H., & Onuchic, J. N. (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298, 937-53, PMCID: 10801360.
31. Koga, N. & 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,

PMCID: 11601854.
32. Frauenfelder, H., Sligar, S. G., & Wolynes, P. G. (1991) The energy landscapes and motions of proteins. *Science* 254, 1598-603, PMCID: 1749933.
33. Sutto, L., Lätzer, J., Hegler, J. A., Ferreiro, D. U., & Wolynes, P. G. (2007) Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc Natl Acad Sci U S A* 104, 19825-30, PMCID: 18077415.
34. Ferreiro, D. U., Hegler, J. A., Komives, E. A., & 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, PMCID: 21273505.
35. Yang, S., Cho, S. S., Levy, Y., Cheung, M. S., Levine, H., Wolynes, P. G., & Onuchic, J. N. (2004) Domain swapping is a consequence of minimal frustration. *Proc Natl Acad Sci U S A* 101, 13786-91, PMCID: 15361578.
36. Ferreiro, D. U., Komives, E. A., & Wolynes, P. G. (2014) Frustration in biomolecules. *Q Rev Biophys* 47, 285-363, PMCID: 25225856.
37. Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* 155, 948-62, PMCID: 24183448.
38. Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113-20, PMCID: 24071849.
39. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., & Campbell, P. J. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43, D805-11, PMCID: 25355519.
40. 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., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S.,

Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., & Stratton, M. R. (2013) Signatures of mutational processes in human cancer. *Nature* 500, 415-21, PMCID: 23945592.

41. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, Jr, L. A., & Kinzler, K. W. (2013) Cancer genome landscapes. *Science* 339, 1546-58, PMCID: 23539594.

42. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. (2004) A census of human cancer genes. *Nat Rev Cancer* 4, 177-83, PMCID: 14993899.

43. Cheng, F., Jia, P., Wang, Q., Lin, C.-C., Li, W.-H., & Zhao, Z. (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol Biol Evol* 31, 2156-69, PMCID: 24881052.

44. Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Di Genova, A., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assunção, J. A., Haggarty, B., Han, D.-J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z.-L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.-H., Letellier, T., Li, C.-

Y., Lio, P., Liu, C.-J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.-J., & Kasprzyk, A. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43, W589-98, PMCID: 25897122.

45. Sali, A. & Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234, 779-815, PMCID: 8254673.

46. Webb, B. & Sali, A. (2014) Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* 47, 5.6.1-5.6.32, PMCID: 25199792.

47. Hubbard, S. J., Campbell, S. F., & Thornton, J. M. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J Mol Biol* 220, 507-30, PMCID: 1856871.

48. Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15, 901-13, PMCID: 15965027.

49. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., & Gerstein, M. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 28, 2267-9, PMCID: 22743228.

50. Ding, L., Wendl, M. C., McMichael, J. F., & Raphael, B. J. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 15, 556-70, PMCID: 25001846.

Page 5: [1] Deleted SK 2/13/16 10:24 PM

Page 5: [2] Deleted SK 2/13/16 10:24 PM

We collected and annotated 6.46 million non-synonymous SNVs using VAT [49]. About 5.1 million of these SNVs were benign mutations that were obtained from the ExAC Project, and an additional roughly 0.6 million SNVs were taken from phase 3 of the 1000 Genomes Project, which constitutes 79% and 9% of our total set of annotated SNVs, respectively (Figure 1b). The remaining SNVs were a set of disease-associated mutations, and these comprised ~76,000 HGMD SNVs and 0.65 million publicly available pan-cancer somatic SNVs. HGMD and the pan-cancer dataset constituted 2% and 10% of the total collected non-synonymous SNVs, respectively (Figure 1b).

However, the contribution of SNVs from different resources changed significantly while considering only those annotated SNVs, which mapped to high-resolution protein structures. Approximately 96,000 SNVs from ExAC were mapped to protein structures in the PDB constituting 51% of our totals set of structurally mapped SNVs (Figure 1c). Similarly, 1KG SNVs constituted 7% (13588) of the total structurally mapped SNV dataset. In contrast, the percentage of the disease-associated SNVs that were mapped to protein structures was 18% (33,261 SNVs) and 24% (44,094 SNVs) for the HGMD and pan-cancer resource, respectively (Figure 1c). The majority of SNVs from the pan-cancer dataset that were mapped to protein structures impacted cancer-associated genes (CAG), constituting 14% (25,409) of the all SNVs mapped to protein structure, whereas SNVs impacting non-cancer associated genes constituted only 8% (15,044) (Figure 1c). In contrast, 4,041 SNVs affecting driver genes mapped to protein structures; these SNVs constitute 2% of the total structurally mapped non-synonymous SNVs (Figure 1c).

Page 9: [3] Deleted SK 2/13/16 10:24 PM

Page 10: [4] Deleted SK 2/13/16 10:24 PM

7).

SNVs that affect oncogenes induce

Page 17: [5] Deleted	SK	2/13/16 10:24 PM