Structural Bioinformatics Enhances Mechanistic Interpretation of Genomic Variation, Demonstrated Through the Analyses of 935 Distinct RAS Family Mutations

Swarndeu Tripathi,1,2 Nikita R. Dsouza,1,2 Raul Urrutia,2,3 Michael T. Zimmermann1,2,4,5,*

1Bioinformatics Research and Development Laboratory, Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA
2Precision Medicine Simulation Unit, Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA
3Department of Surgery, Medical College of Wisconsin, Milwaukee, WI 53226, USA
4Clinical and Translational Sciences Institute, Medical College of Wisconsin, Milwaukee, WI 53226, USA
5Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI 53226, USA

*To whom correspondence should be addressed.

Abstract
Motivation: Protein-coding genetic alterations are frequently observed in Clinical Genetics, but the high yield of variants of uncertain significance (VUS) remains a limitation in decision making. RAS-family GTPases are cancer drivers, but only 54 variants, across all family members, fall within well-known hotspots. However, extensive sequencing has identified 881 non-hotspot variants for which significance remains to be investigated.

Results: Here, we evaluate 935 missense variants from seven RAS genes, observed in cancer, RASopathies, and the healthy adult population. We characterized hotspot variants, previously studied experimentally, using 63 sequence- and 3D structure-based scores, chosen by their breadth of biophysical properties. Applying scores that display best correlation with experimental measures, we report new valuable mechanistic inferences for both hot-spot and non-hotspot variants. Moreover, we demonstrate that 3D scores have little-to-no correlation with those based on DNA sequence, which are commonly used in Clinical Genetics. Thus, combined, these new knowledge bear significant relevance.

Contact: mtzimmermann@mcw.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction
Genetic variants activating Rat Sarcoma (RAS) genes are among the most recurrent somatic alterations in human cancers, affecting up to 25% of solid tumors. (Hobbs, et al., 2016) Their somatic hotspots have been extensively studied, yet mechanistic understanding for the experimentally measured differences among certain hotspot variants is lacking, and non-hotspot variation, while collectively common, has not been studied. RAS is thus a protein of high biomedical importance that also highlights a current challenge in translational genomics, which is the functional interpretation of mutations. (Andreoletti, et al., 2019; Hu, et al., 2019) We
designed and describe herein a systematic approach to scoring the effects of protein-coding genomic variants by accounting for a wide array of sequence- and structure-based effects, in order to identify mechanistic and testable hypotheses for the effects of hotspot and non-hotspot RAS variants. Therefore, the current study focuses on RAS as a demonstration of the scalability of our approach and the importance to integrate a broad range of computational biophysical scores because each provides unique information for functional interpretation.

To interpret genomic variants, current clinical genomics guidelines rely on recurrent observations in disease cases or tumors, compared to the general healthy population, and inferred impact on the encoded protein. (Karbassi, et al., 2016; Richards, et al., 2015) However, in our view, the gene product itself must take center stage, either using bioinformatics, functional validation, or both. The fundamental concept underlying this idea is that protein structure and dynamics play a key role in determining whether a missense variation can be tolerated or becomes pathogenic. Further, current approaches aim to directly predict pathogenicity resulting in different levels of predictive performance. (Hart, et al., 2019; Karchin, et al., 2007; Ponzoni and Bahar, 2018) We believe that this type of operation bypasses the necessary step of determining the molecular mechanism of dysfunction. Thus, the wide adoption of protein 3D structural biology is of paramount importance since the mechanistic interpretation of novel genomic variants will ultimately be inferred from the study of the gene product.

The RAS family of small GTPases that cycle between a guanosine triphosphate (GTP)-bound (active) and guanosine diphosphate (GDP)-bound (inactive) form, (Milburn, et al., 1990) has 31 members and all act as signal transducers influencing cellular growth and differentiation. Genetic variants in RAS proteins, when present in the germline, are responsible for rare congenital diseases known as RASopathies, (Grant, et al., 2018; Simanshu, et al., 2017; Tidyman and Rauen, 2009) such as Noonan (RRAS) and Costello (HRAS) Syndromes. The most clinically relevant variants, due to their common oncogenic mutation, are KRAS, HRAS, and NRAS (Rauen, 2013), discovered from the study of oncogenic viruses and neuroblastomas. (Cox and Der, 2010) Additional family members have been identified through sequencing of tumors (MRAS), RASopathy patients, and neural transformation (e.g. RERG and RRAS2).

In cancer, there are two highly recurrent RAS activating variant sites, referred to as hotspots, but many genetic alterations are observed outside of the hotspot sites somatically in cancer, in RASopathies, and in the currently healthy adult population. This genetic spectrum differs for each member of the RAS family. A growing body of experimental data indicates that each type of alteration at the hotspots can lead to different downstream effects including different GTPase activities, nucleotide exchange rates, effector preference, cellular morphologies, and neoplastic potential. (Angeles, et al., 2019; Bandaru, et al., 2017; Burd, et al., 2014; Cirstea, et al., 2013; Ilihe, et al., 2012; Munoz-Maldonado, et al., 2019; Seeburg, et al., 1984; Smith, et al., 2013) However, hotspot variants have not been uniformly assessed by the same experimental assays, making interpretation of many hotspot variants uncertain. Further, non-hotspot variants may not alter the protein in the same way as hotspot mutations and therefore may not have the same implications for clinical management. Thus, better methods to evaluate how genetic variation affects RAS, within and outside of the hotspots, are needed in order to interpret their potential functional effects.

Recurrent cancer variants such as KRAS G12D and G12V have been extensively studied by laboratory experimental and computational methods (Ioanidis, et al., 2016) and found to be distinct from other RAS hotspot variants, even at the same codon. (Burd, et al., 2014; Munoz-Maldonado, et al., 2019) However, the full spectrum of rare disease and cancer variants has not been evaluated with the same rigor. Therefore, the goal of this study is to help fill the gap in knowledge by evaluating a more comprehensive series of 3D scores, integrated with DNA and protein sequence-based scores (Figure 1), for interpreting the most likely underlying mechanisms of hotspot alteration by genetic variants in members of the RAS family of proteins, and assessing the potential for non-hotspot variants to have similar effects. Our results show that no single score is likely to capture enough detail to fully interpret the effects of RAS genomic variants. Rather, different scores indicate alterations of specific functional properties that are not available from DNA annotations, especially among 3D protein structure-based scores. Combined, these results demonstrate that when carefully parametrized, 3D scoring methods from structural bioinformatics are superior in mechanistic information than the conventional DNA-based which are currently widely used in Clinical Genetics. Thus, this new information extends not only our understanding of RAS proteins, which is of paramount medical importance but also increases the arsenal of analytic methods available to medical genetics.

2 Methods

2.1. Selection of RAS Family Proteins and Molecular Modeling

We surveyed the PDB (Berman, et al., 2000) for experimental structures of the most commonly altered RAS family proteins in cancer and germline RASopathies. We identified that for RRAS, RRAS2 and RERG only a single experimental structure was available, which is the GDP-bound structure. Thus, for consistency, GDP-bound structures were used for all the RAS proteins in this study (Table S1). We included the three most altered RAS proteins in cancer, and four additional RAS members that are most known in RASopathies. To model the missing loops in NRAS (residue 61–71) and RRAS2 (residue 71–74 in chain A) we used the ModLoop web server for automated modeling of loops in protein structures. (Fiser, et al., 2000; Fiser and Sali, 2003) The mapping between amino acids of the seven RAS-family proteins used in this study was defined by their protein sequence based on multiple sequence alignment (MSA), shown in Figure S1.

We included the GTP-bound or GppNHp (non hydrolyzable GTP analog)-bound RAS structures for KRAS, HRAS, and NRAS in an analysis of the sensitivity of structure-based scores to the input 3D conformation. As a proof of concept, we compared 3D scores between the WT GDP and GppNHp-bound 3D structures of KRAS, HRAS and NRAS (Table S1). We found strong correlations among the 3D scores for the GDP and GppNHp bound structures. Nonetheless, some local differences were observed at the residue level (see the Results section 3.4). However, such differences are a key advantage of structure-based scores, as well as a challenge since context matters. Thus, our approach is based on a defined context and mechanistic conclusions should be interpreted appropriately.

2.2. Genomic Variant Annotation

We defined the GTPase domains of seven RAS family proteins in 3D and identified their corresponding DNA coding regions in the human genome (GRCh38) for UniProt isoforms (Table S1 and Figure S1). We used COSMIC (Forbes, et al., 2011) to identify genetic variants previously observed somatically in human cancers. We used ClinVar (Landrum, et al., 2014) and HGMD (Stenson, et al., 2017) to identify genomic variants responsible for congenital diseases and RASopathies. We identified variants observed in the currently healthy adult population, and their corresponding global minor allele frequency (MAF), from gnomAD (Karczewski, et al., 2019). Annotations for DNA sequence-based scores were gathered from dbNSFP v4.0 (Liu, et al., 2011). We used the BioR
Correlation among all 63 scores that we selected for consideration in our combined approach are shown in Figure S2A. Using their mutual correlations, we selected 31 scores that were largely distinct representatives from the 63 (Figure S2B).

Analysis using t-distributed stochastic neighbor embedding (t-SNE) construct a low-dimensional 2D embedding of high-dimensional data using the distances between variants and using our combined set of computational scores. We used the Rtsne package v0.15 (Krijthe, et al., 2018) based on optimized threshold estimate for a trade-off between speed and accuracy in a Barnes-Hut implementation of t-SNE using Euclidian distances, (Van Der Maaten, 2014) a perplexity factor of 80, and theta set to 0.3. We also combined scores for comparing among variants using PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding), which is another newly developed dimensionality reduction technique. PHATE generates a low-dimensional embedding, in a way that attempts to preserve local and global similarities, generating clusters and branches within the data, to enhance interpretability of the data’s underlying structure, (Moon, et al., 2019) which in this study is the similarities among different types of computational scores’ values for RAS variants. We generated heatmap plots to identify the patterns of the computational scores among the hotspot variants using the pheatmap v1.0.12 package (Kolde and Kolde, 2015) and the complete hierarchical approach using Euclidean distances after scaling each score to z-scores.

3 Results

We aim to better understand molecular mechanisms that underlie alterations caused by genetic variation in RAS members, at hotspots but, more importantly, for the entire landscape of missense human variation in RAS identified to date, defined by integrating across leading databases in clinical genetics, cancer, and the healthy population (see Materials and methods). To draw comparison with methods currently used by genomic analysts, we computed distinct scores that account for data from DNA and protein sequence, as well as 3D protein structural levels (Figure 1A). We also assessed the available experimental data for RAS hotspot variants (Figure 1B and Table S3). Thus, using those well-studied variants, we identified computational scores that have the strongest associations with damaging effect on the encoded protein, as surrogates for their underlying mechanisms of dysfunction (Figure 1C). Then, we assess all 935 variants observed in RAS, to quantify the relationships among the scores, as well as to score the non-hotspot variants.

Our targeted analyses involved 63 selected scores that capture information present in genomic DNA annotations, protein sequence properties, and protein structural features (12 DNA sequence, 9 protein sequence, and 42 3D structure-based scores, respectively) (Figures 1D and S2A). These features were manually selected from many hundreds of candidate features using domain knowledge and literature precedence for best-in-class and unique measures of protein properties. Next, we examined the correlation structure to further refine the set of granular scores. First, by choosing a representative from pairs of scores that have a low absolute correlation (Rgranular ≤ 0.4). Second, we included 2 frustration-based and 1 multi-body potential scores for their assessment of unique biophysical properties that therefore most efficiently cover the broadest diversity of RAS properties. Third, we included SIFT and CADD because of their common use in the field; SIFT is highly correlated with CADD (+0.66) and Evolutionary Trace (ET) (0.57) (Mihalek, et al., 2004) and CADD is highly correlated with DANN (0.48) and ET (-0.44). This procedure resulted in a final filtered selection of 31 granular scores (Figure S2B). To guide our mechanistic characterization of the variants, we compared
these 31 scores with direct experimental measurements, such as intrinsic GTP hydrolysis rates ($K_{\text{on hydrolysis}}$) of the eight G12, G13 and Q61 hotspot variants (Figure 1E). Noteworthy, we find that this method of experiment-guided parameterization of the computational scores for the hotspot variants (Figure 1E) facilitates mechanistic interpretation derived for the non-hotspot variants. This finding is important since, although RAS hotspot variants are highly studied, they have been assessed only in distinct RAS members and using different experimental assays, making direct comparisons between them challenging (Table S3). Further, we find that the classic paradigm that RAS hotspot mutations are activating, derived from early studies that assessed the ability of variants to activate neoepitopic potential, (Bos, 1989; Corominas, et al., 1991; Kumar, et al., 1990) is clearly not the case for all variants, from an enzymatic perspective. We chose the study of Hunter et al. (Hunter, et al., 2015) as our experimental measures for parameterizing structural bioinformatics scores because they evaluated a useful set of KRAS hotspot variants using quantitative assays. Therefore, these experimental data provide a useful tool for parameterizing our search for mechanistic associations via an integrated protein scoring procedure.

Figure 1: Our process for integrating computational scores from multiple molecular levels with experimental data to interpret hotspot molecular mechanisms of genomic variants. A) Multiple distinct molecules carry relevant information for directly interpreting the effects of genomic variants: the DNA itself, the encoded mRNA, the linear protein, and the 3D folded protein. B) Schematic of GTP hydrolysis kinetics and nucleotide exchange kinetics are shown for RAS GTase. C) Our long-term goal is to predict altered mechanisms. In this study we take the first step of aggregating multiple and diverse scores across molecular levels and correlating them with experimental measures of activity. D) We have assembled 63 computational scores for how genomic variants may alter sequence or structure and analyzed their interrelationships, with the number of scores from each molecular level (DNA sequence, protein sequence, and 3D structure) shown in parentheses. E) Measurements of intrinsic GTP hydrolysis rate of RAS hotspot variants shown as a heatmap in descending order indicating low and high rates relative to the WT RAS (left panel) (Hunter, et al., 2015) Assessment of underlying molecular mechanism of RAS hotspot variants shown as a heatmap by correlating experimental measurements (or scores) and computational scores (right panel) for Spearman correlation, $R_{\text{spearman}}$ (Table S5) indicated below the dendrogram. The heatmap colors correspond to the z-scores, high (red) and low (blue) while blue and light red dendrogram at the top of the heatmap represent positive and negative $R_{\text{spearman}}$ respectively.

3.1. Combined Scores Explain Mechanisms Underlying Hotspot Experimental Findings

For explaining mechanisms by which hotspot variants disrupt RAS function, we first employed four quantitative experimental measurements (intrinsic hydrolysis rate ($K_{\text{on hydrolysis}}$), GAP-stimulated $K_{\text{on hydrolysis}}$, GDP and GTP exchange rates, and RAF affinity (Hunter, et al., 2015) for eight KRAS variants (G12C, G12D, G12A, G12V, G12R, G13D, Q61L and Q61H), and compared to WT (Table S4). Next, we correlated these measurements with our multi-tier computational scores (Figure S3 and Table S5). We identified seven computational scores (one DNA sequence, one protein sequence and five 3D-structured based scores) that strongly correlate ($R_{\text{spearman}} \geq 0.6$) with the intrinsic $K_{\text{on hydrolysis}}$ values (Table S5 and Figure S4A). Next, we represented these scores as a heatmap to better visualize their relationship with intrinsic $K_{\text{on hydrolysis}}$ (Figure 1E). In this manner, we grouped the KRAS hotspot variants, one with high (G12C, G12D and G13D) and another with low (G12V, G12R, G12A, Q61H and Q61L) intrinsic $K_{\text{on hydrolysis}}$. When carefully analyzed, we find that the pattern of computational scores further highlights the mechanistic differences among hotspot variants. Importantly, we identified that changes in 3D scores representing average frustration, configurational energy, and main-chain solvent accessible surface area, and the protein sequence-based score quantifying the change in electron-ion interaction potential, have the strongest association with impaired intrinsic $K_{\text{on hydrolysis}}$ by hotspot mutations ($p \leq 0.05$) (Table S5). Therefore, we propose that these scores capture specific contributions to the underlying mechanism of altered RAS function. Knowing that RAS proteins bind critical effectors, we investigated their association with experimental measures of RAF affinity and identified six 3D structure-based scores that strongly correlate ($R_{\text{spearman}} \geq 0.5$; Figures S4A and B). These scores indicate that changes in local unfolding propensity, structural stability, and hydrophobic solvation energy, appear to be important contributors to the mechanism of impaired RAF affinity by hotspot mutations. Similar comparisons for the further two experimental measures of GAP-stimulated $K_{\text{on hydrolysis}}$, and nucleotide exchange rates are shown in Figures S4C and D, respectively.

Therefore, using multiple experimental measures as benchmarks and integration of multiple structural scores, our analysis is able to identify mechanistic scores that differentiate each hotspot variants’ effect at the molecular level.

3.2. Experiment-Guided Scores for Hotspot Variants Applied to Non-Hotspot Variants for Mechanistic Interpretation

Due to the limited availability of systematic and quantitative experimental measurements of RAS variants in the literature, supervised machine learning approach is not feasible in this study. As a result, we specifically focused on the KRAS hotspot variants for which experimental measurements are available using unsupervised machine learning methods. Patterns among experimental measurements demonstrate the unique profile of each variant, and which computational scores most closely associate with each (Figure S4). We extended our mechanism investigation for all RAS variants using an unsupervised dimensionality reduction method, PHATE (see Materials and methods), that captures both local and global patterns using an information-geometric distance. (Moon, et al., 2019) PHATE captures the similarity between nearby data points using nonlinear transformation by converting the Euclidean distances into ‘local affinities’ and preserves the global similarities between data points using data diffusion. Therefore, PHATE is especially suitable for mechanism characterization of RAS hotspot vs. non-hotspot variants. First, we focused our 2D PHATE analysis using the computational scores that strongly correlates with the experimental data of intrinsic $K_{\text{on hydrolysis}}$ from the previous section (Figure 1E and Figure S4C). In particular, we analyzed all the 935 variants from 7 RAS proteins and highlighted those 8 KRAS hotspot variants for which intrinsic $K_{\text{on hydrolysis}}$ data is available (Figure 2A). Additionally, showing 493 (54
hotspots and 439 non-hotspots) variants from the three most recognized proto-oncogenic RAS (KRAS, HRAS and NRAS) proteins separately, we again highlighted those same KRAS hotspot variants (Figure 2B). We note that in the 2D PHATE space G12, G13 and Q61 variants are distinct in Figures 2A and B further elucidating different mechanism impacting intrinsic \( k_{\text{hydrolysis}} \) at these hotspot sites. Interestingly, differences in \( k_{\text{hydrolysis}} \) among KRAS G12 variants are evident, such as between G12A/R and G12C/D/V with lower and higher hydrolysis rates, respectively. Though these variants are overall nearby each other in the 2D PHATE space (Figure 2A and B), our interpretation is that the G12 position has a specific effect on KRAS structure when altered, and computational scores identify this overall effect in addition to how certain variants may have distinct effects from one another at the same site. Next, we performed similar 2D PHATE analyses utilizing the computational scores that strongly correlate with the GAP-stimulated \( k_{\text{hydrolysis}} \) (see Figure S4C) for all the 935 (from 7 RAS) and 493 (from HRAS, KRAS and NRAS) variants (Figures 2C and D, respectively). Unlike in the 2D PHATE space for \( k_{\text{hydrolysis}} \), the hotspot variants in the 2D PHATE space for GAP-stimulated \( k_{\text{hydrolysis}} \) are more distinctly separated by the magnitude of the impact relative to the WT GAP-stimulated \( k_{\text{hydrolysis}} \). We identified similar patterns in the data for RAF affinity and GDP/GTP exchange rates by PHATE analyses (Figure S5). Therefore, we believe diverse computational scores across all three molecular levels can be used to enhance the interpretation of position- and variant-specific mechanistic effects.

![Figure 2: KRAS non-hotspot variants computationally prioritized for effects on intrinsic and GAP-stimulated hydrolysis rates. We used the correlated computational scores (see Table S4) to assess all non-hotspot variants for their potential to alter intrinsic and GAP-stimulated \( k_{\text{hydrolysis}} \) respectively similar to hotspot variants. Because we are specifically interested in global patterns among the variants, we used PHATE for dimensionality reduction. A-B) 2D PHATE analysis was performed on 935 variants from 7 RAS in (A) and 493 variants from HRAS, KRAS and NRAS in (B) consisting the five computational scores that correlate with the intrinsic \( k_{\text{hydrolysis}} \) of the KRAS hotspot variants (Figure 1E). Eight hotspot somatic variants of KRAS are colored based on the intrinsic \( k_{\text{hydrolysis}} \) measurements relative to the WT in the 2D PHATE plots in both (A) and (B). C-D) Similar to (A) and (B) 2D PHATE analysis was performed on 935 (from 7 RAS) and 493 (from HRAS, KRAS and NRAS) variants in (C) and (D), respectively using the seven computational scores that correlate with the GAP-stimulated \( k_{\text{hydrolysis}} \) (Figure S4C). In both (C) and (D) eight hotspot somatic variants of KRAS are colored based on the GAP-stimulated \( k_{\text{hydrolysis}} \) measurements relative to the WT. High (red) and low (blue) hydrolysis rates are indicated in the color bar. E) 3D structure of KRAS (PDB: 40BE) showing the sensitive regions, phosphate-binding loop (P-loop) (amino-acid 16-17), switch-I (amino-acid 30-40) and switch-II (amino-acid 60-76). (Johnson, et al., 2017) F) Amino-acid residues of the KRAS variants that are nearby to the eight hotspot variants in the 2D PHATE space in (B) are shown projected onto the 3D structure. The amino-acid residues are colored according to the sensitive regions in (E).](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa972/5998662)
and G12D, which tightly clustered with each other and across RAS proteins, while G12C, which was distinct from them also tightly clustered across RAS proteins (Figure 3A). In addition, amino acid substitutions for alanine, serine, or arginine at G12 and G13 are distinct from other hotspot variants, but more variable between RAS proteins and exhibited greater differences between G12 and G13 sites (Figures 3A-B). We further noticed that Q61 variants (Figure 3C) are more widely distributed in the t-SNE space compared to the G12 and G13 variants. This is especially true for the arginine, proline, and glutamic acid substitution at Q61. From these patterns among hotspots variants, we hypothesize that certain hotspot variants affect the same molecular mechanism across all RAS proteins, while others have a more distinct mechanism.

Figure 3: All RAS variants assessed using our integrated scoring approach identify functional groupings among hotspot variants and demonstrate differences among non-hotspot variants. Having assessed global patterns by PHATE, we next characterized more nuanced local patterns among the variants using 2D t-SNE. Combinations of scores convey mechanistic differences in the effect of different hotspot variants. A-C) Somatic hotspot variants of G12 (A), G13 (B) and Q61 (C) from HRAS, KRAS and NRAS are shown in the 2D t-SNE space of all the 7 RAS variants. Hotspot variants are labeled and colored by RAS protein. It is visually apparent that some non-hotspot variants are nearby hotspot variants in t-SNE space, indicating that they may have similar effects as hotspot variants, while other non-hotspot variants are far from hotspot variants in the t-SNE space, indicating that they either have no effect or a different effect from hotspot variants. D) Heatmap plot shows patterns of scores for the somatic hotspot variants shown in the 2D t-SNE plots. The top ten scores (out of 31) are selected based on median absolute deviation (MAD ± 0.2). We separated the G12, G13 and Q61 variants in five clusters in the heatmap plot denoted as (i)-(v). Heatmap plot with all the 31 scores are shown in Figure S6.

Next, we identified the most variable scores across genetic variants (median absolute deviation ± 0.2), consisting of one DNA sequence, three protein sequence, and six 3D structure-based scores, for the G12, G13 and Q61 variants in KRAS, HRAS and NRAS. We visualized the ten most variable scores to identify patterns among hotspot variants for each RAS protein and from different RAS proteins (Figure 3D). We also represented all the 31 scores as a heatmap for the G12, G13 and Q61 variants in KRAS, HRAS and NRAS (Figure S6). This approach let us identify global patterns that distinguish effects among G12, G13 and Q61 variants, which were primarily driven by changes in mainchain entropy and structural stability (ΔΔGheat). These two structure-based metrics indicate a different mechanism for G12, G13 and Q61 variants whereby G13 variants have a stronger effect on stability (shown inside a rectangle in Figure 3D). Especially, the G13 variants are highly destabilizing (Median ± MAD; ΔΔGheat=−3.5 ± 0.6 kcal mol$^{-1}$), whereas G12 variants are either stabilizing or neutral (ΔΔGheat=0.15 ± 0.22 kcal mol$^{-1}$). On the other hand, Q61 variants are moderately destabilizing (ΔΔGheat=−0.37 ± 0.07 kcal mol$^{-1}$). In addition, we identified RAS-specific clusters that display differences in Evolutionary Trace (ET) scores for sequence conservation for G12 and G13 variants, but interestingly similar differences in ET scores are not observed for the RAS-specific Q61 variants. Moreover, we note that irrespective of the RAS gene all G12C variants are clustered together (indicated as cluster ‘(i)’ in Figure 3D), consistent with t-SNE visualization, indicating that the alternate amino acid has similar effect in all three proteins. Note that G13C KRAS and G13C NRAS cluster together (cluster ‘(ii)’ in Figure 3D) and show similar pattern in their scores, while G13C HRAS clusters with the other three HRAS G13 variants (G13A, G13D, and G13S) (cluster ‘(ii)’ in Figure 3D). Next, we considered the patterns among the five other variants (G12A, G12D, G12R, G12S, and G12V) at G12. These five G12 variants showed a distinct pattern compared to G12C and clustered together (cluster ‘(ii)’ in Figure 3D) for each RAS protein. Within these RAS-specific G12 clusters, we identified that these five variants affect HRAS differently than NRAS and KRAS, by displaying a markedly higher change in 3D contact-level frustration, which is likely supported by a more energetically favorable change for GDP-bound HRAS than for GDP-bound NRAS or KRAS. When considering changes at the G13 position, we find patterns among the scores that distinguish HRAS in particularly from NRAS and KRAS (clusters ‘(iii)’ in Figure 3D). This finding indicates that the context given to G13 by the HRAS intrinsic protein environment may differ for G13 compared to KRAS or NRAS and is primarily driven by differences in solvent accessible surface area and conformational frustration, similar to G12 differences and the CADD scores. For Q61 we identified all the variants from KRAS, HRAS and NRAS in cluster ‘iv’ except Q61L from NRAS, which was identified as an independent cluster ‘v’). Interestingly, the cluster ‘iv’ variants are all at Q61 and have a distinct pattern reflecting changes in heat capacity and pK_a(RCOOH). Thus, together, these results reveal distinct patterns of scores for the G12, G13 and Q61 hotspot RAS variants, indicating that they may have distinct functional effects, congruent with growing experimental evidence. Our findings not only elucidate the sensitivity of structure-based scores but also demonstrate the potential gains by integrating them with DNA sequence-based scores for interpreting the impact of genomic variants on RAS function.

3.4. Sequence- and Structure-Based Scores Provide Distinct Information From One Another

In the analyses discussed above, we present a comparative analysis of both sequence-based and structure-based scores for all 935 unique observed missense RAS variants (Figure 1A), with the methodologic goal of systematically assessing their individual and combined utility for interpreting genomic variants. Thus, we next focused on the information content among the scores, as applied to changes due to all RAS variants. For this purpose, we explored the correlation among scores from different molecular levels. We chose three illustrative examples of negative, neutral, and positive associations among scores from different molecular levels. First, we detect that differences in protein sequence-based heat capacity is negatively correlated with differences in structure-based
hydrophobic solvation free energy (Figure 4A). This finding shows that increased (decreased) heat capacity of a variant associates with favorable (unfavorable) changes in hydrophobic solvation free energy. Second, interestingly, changes in local frustration index show no correlation with the change in residue sidechain folding cooperativity (Figure 4B) even though both are structure-based scores, indicating that these scores report on two independent properties. Third, we identified positive correlation between the DNA and protein sequence-based SIFT and ET scores (Figure 4C), which is expected since conserved residues are more likely deleterious than variants from non-conserved residues and SIFT leverages conservation. Thus, scores between molecular levels may be related to each other, but more importantly, they can contain information that is not apparent at other molecular levels.

**Figure 4:** Correlation among computational scores from multiple molecular levels for the 935 RAS variants demonstrates the distinct and underutilized value of 3D structure-based scores. A-C) Spearman correlation ($R_{Spearman}$) among pair of computational scores indicating negative (A), neutral (B) and positive correlation (C). These three examples were chosen as exemplars for relationships between the scores from different molecular levels. D) We used total 65 individual scores for the 935 protein variants to assess the differences among variants from 7 RAS proteins (KRAS, HRAS, NRAS, MRA, RAS, RRAS2 and RERG) based on DNA sequence, protein sequence and 3D structure of protein. Larger and labeled versions of the correlation matrices are available in Figures S2. We highlight two sections of structure-based scores that have nearly no overlap with one another or with information available in DNA annotations. E) Using correlation patterns among scores, we reduced the number of individual scores to the 31 that are most unique and therefore most efficiently cover the broadest diversity of properties. The locations of same three scores specifically named in (D) are indicated by arrows. In (D) and (E) the size of each small square in the correlation matrix is proportional to the value of absolute Spearman correlation ($|R_{Spearman}$). All 63 scores for 935 variants from 7 RAS genes are provided as Supporting Data (Table S2).

Subsequently, we computed the correlations among all 63 computational scores from three molecular levels (Figures 4D and S2A). We found that protein structure-based scores have little-to-no correlation with the DNA sequence-based scores, overall (also see Figure S7), indicating that they represent an entirely different type of information about the effects of genomic variants. Further, among structure-based scores, there is a high diversity with some scores having little-to-no correlation with each other because they assay different protein physicochemical characteristics. This result indicates that, as performed above, a more integrated approach is required to better interpret genomic variants. We found that 3D scores used in this study, are robust to the input experimental structure (Figure S8A) with mean $R_{Spearman} = 0.92$ (Figure S8B). They also demonstrate high specificity (Figure S8C). High specificity is an advantageous feature, demonstrating the biophysical and mechanical detail of 3D scores – residues that are experimentally known to differ between the GDP and GTP bound forms are those with largest differences in 3D scores (Figure S8C). Finally, we selected a subset of 31 scores based on the correlation structure among scores (when individually applied to all variants), and their assessment of different biophysical or biochemical properties, so that they are unique and most efficiently cover the broadest diversity of RAS properties (Figures 4E and S2B). Thus, we propose that this set of scores represents an efficient coverage of many unique dimensions of protein function and should serve as a baseline for developing a more comprehensive system for interpreting the effects of genomic variants.

In summary, DNA sequence-based information is the mainstay of genomics data interpretation, but we have demonstrated in this work that additional information relevant for interpreting genomic variants and not currently available from the DNA sequence, can be derived from computational study of the protein 3D structure. RAS is a critically important proto-oncogene with a broad spectrum of non-hotspot variation that we have assessed for similarity to experimental enzymatic properties of hotspot mutations. There are multiple categories of 3D structural features that indicate different properties of RAS proteins altered by genomic variants, and likely of proteins in general. Therefore, assaying one type of feature is insufficient for genomics data interpretation. This study demonstrates the potential that an integrated approach provides. A full suite of scores that integrate across all biologic layers of molecular function is required and must be considered in future improvements to the guidelines for genomic data interpretation.

4 Discussion

Currently, there are no 3D structure-based methods that have been standardized and parameterized for use in clinical genomics workflows, yet there is high potential for them to add value to the interpretation of genomic variants and indicate altered molecular mechanisms. To test this potential, we assessed KRAS hotspot variants that have been experimentally measured (Figure 1E) and showed that mechanistic inferences can be made for why some variants alter enzymatic properties and others do not (Figure 2). We then applied the computational scores that associate with enzymatic function to non-hotspot variants to predict which of them may have similar effects to hotspot mutation (Figure 3). Next, we assessed a wide range of structure-based scores and applied them to all 935 RAS hotspot and non-hotspot variants (Figures 4D and E), demonstrating first the significant added information from structure-based scores, and second the ability to scale these methods to large numbers of variants. We believe that the current paradigm in the field, of aiming to directly predict the pathogenicity of variants, skips the critical step of inferring, with precision, molecular mechanisms of dysfunction. This study integrates the broadest diversity of scores (by type and nature of the scores) to date for mechanistic characterization of RAS non-hotspot variants and provides a scalable framework for application to other proteins with variants identified by high-throughput sequencing.

We have assembled a large and diverse group of scores from DNA annotations, protein sequence properties, and protein 3D properties and used it to show that 3D-scores enhances the information available from the DNA, leading to greater specificity. As a first example, several studies have shown that pair-wise energetic potentials can reliably identify changes in $\Delta\Delta G_{m,s}$ associated with genomic variants. (Cheng, et al., 2006; De Baets, et al., 2011; Yang, et al., 2013) Second, compared to pair-wise
potentials, four-body contact potentials were developed to identify the best 3D models from a set of candidate models because they better capture the non-linear protein fold as well as the many interactions between the protein backbone, side chains, and solvent. (Feng, et al., 2007) Thirdly, due to allosteric transitions and biomolecular interaction sites, variants throughout the protein can lead to local functional changes without being destabilizing. In this context, local frustration quantifies the balance (or imbalance) among energetically favorable and unfavorable interactions (Ferreiro, et al., 2007) and has been shown useful to interpreting the impact of genomic variants. (Kumar, et al., 2016) We further quantified an experimentally parameterized thermodynamic measure of local foldability using transfer free energies based on residue-specific implementation of the SEED algorithm, (Porter and Rose, 2012) which parses proteins into their constituent thermodynamically cooperative components. (Zimmermann, et al., 2015) Finally, we integrated scores and generated topologic groups that we believe may indicate different molecular mechanisms and thereby probabilities of pathogenicity. Previous studies have shown the importance of protein 3D structure (Berliner, et al., 2014; Dixit, et al., 2009; Karchin, et al., 2007; Kiel and Serrano, 2014) and dynamics (Dixit and Verkhivker, 2014; Ponzoni and Bahar, 2018) in assessing functional impact of missense variants, in general. Further, most of the scores we combined have been individually tested against genomic predictors or disease classification. However, no study, to our knowledge, has combined the broad diversity of scores together, with DNA annotations, for the interpretation of genomic variants. By combining them, we identified that many of them are unique, potentially assaying a different dimension of the protein and explaining why they have modest performance on an individual basis. We also assessed if transcripts from the seven RAS-gamily genes were of similar complexities to one another or if they were more diverse and found that they span much of the proteome’s transcript-level local complexity (Figure S9). Thus, our approach is innovative, likely generalizable to other proteins, and has a high potential to elucidate altered mechanisms. The most well studied RAS mechanism of dysregulation is activating hotspot mutation, which is commonly observed in human cancers. A simple functional hypothesis is that all hotspot variants are damaging to function and activating, but an increasing array of evidence is indicating that different changes at the same hotspot residue result in different amounts of dysregulation or even different types of activation. (Prior, et al., 2012) For example, in our data, G12C is more alike across RAS proteins than other G12 variants; our approach indicates that G12C affects RAS differently than other G12 variants. Recent experimental studies have shown that G12C is unique among G12 variants for its ability to be specifically inhibited. (Lindsay and Blackhall, 2019) Functional genomics experiments will be critical for completing our understanding of how mechanistic changes to RAS lead to distinct cellular effects. Concordance between existing experiments and 3D scores highlights the potential utility of our approach for identifying underlying mechanisms.

The primary aims of the current study were to quantify the difference in information content among DNA and 3D structure-based scores, and to investigate if there were groups of RAS variants based on how they alter the landscape of scores. We aim to define new scores that more directly assess biologic mechanisms of dysfunction. That is, to define energetic, parametric, or molecular mechanic scores that conveys an underlying biophysical landscape or mechanism of alteration, even if they do not directly measure that landscape. Our long-term aim is to categorize variants into mechanistic groups. Such mechanisms are the underpinnings for disease. Thus, we aim to predict pathogenicity by first determining mechanism.

We have established our approach to protein structure-based scores as an initiating point for a fuller description for how genetic variants may affect protein function. Pilot studies on RAS proteins, (Clausen, et al., 2015; Gorfe, et al., 2008; Grant, et al., 2009; Ioannidis, et al., 2016) and our own studies on other proteins, (Blackburn, et al., 2017; Klee and Zimmermann, 2019; Long, et al., 2016; Zimmermann, et al., 2018) have demonstrated the utility of atomic molecular simulations to provide additional information such as allosteric transitions and functional motion, but scalable approaches using these tools remain to be developed. Also, more quantitative groupings of the variants, generated by training against a larger amount of experimental functional assays, are required for a more definitive assessment. Our ongoing work will extend the structure-based approach presented here, to include dynamics-based scores, as well as scores derived from protein shape and surface properties. We will extend our application of protein scores to include the GTPassE fold in general and further details for variants determining RASopathies. We firmly believe that the approach we have presented here is applicable to a broad range of the human proteome and will become an important criterion in future versions of guidelines for the interpretation of genomic variants.

Acknowledgements

This research was completed in part with computational resources and technical support provided by the Research Computing Center at the Medical College of Wisconsin.

Funding

This work was supported by National Institutes of Health Grant R01 DK52913 (to R. U.), the Advancing a Healthier Wisconsin Endowment (R. U.) and an endowment from the Linda T. and John A. Mellowes Foundation (R. U.).

Conflict of Interest: none declared.

References

Andreselli, G., et al. Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. *Hum Mutat* 2019;40(9):1197-1201.

Angela, A.K.J., et al. Phenotypic characterization of the novel, non-hotspot oncogenic KRAS mutants E31D and E65K. Oncol Lett 2019;18(4):420-432.

Bandara, P., et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. *Elife* 2017;6.

Berliner, N., et al. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One* 2014;9(9):e107353.

Berman, H.M., et al. The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):233-242.

Blackburn, P.R., et al. A Novel Kleefstra Syndrome-associated Variant That Affects the Conserved TPLX Motif within the Arykin Repeat of EHMT1 Leads to Abnormal Protein Folding. *J Biol Chem* 2017;292(9):3866-3876.

Bos, J.L. ras oncogenes in human cancer: a review. *Cancer Res* 1989;49(17):4682-4689.

Burd, C.E., et al. Mutation-specific RAS oncogenicity explains NRAS codon 61 selection in melanoma. *Cancer Discov* 2014;4(12):1418-1429.

Cheng, J., Randall, A. and Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 2006;62(4):1125-1132.

Cirstea, I.C., et al. Diverging gain-of-function mechanisms of two novel KRAS mutations associated with Noonan and cardio-facio-cutaneous syndromes. *Hum Mol Genet* 2013;22(2):262-270.

Clausen, R., et al. Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Comput Biol* 2015;11(9):e1004470.

Corominas, M., et al. ras activation in human tumors and in animal model systems. *Environ Health Perspect* 1991;93:19-25.
De novo RRAGC mutation activates mTORC1 signaling in 2014;42(Database
Visualizing structure and transitions in high-dimensional The small GTPases K-Ras, N-Ras, and H-Ras have distinct Effect of KRAS oncogene substitutions on protein behavior: Structure-and-signatures-of-cancer-mutation-hotspots-in protein kinases. Comput Math Methods Med 2014;2014:633487. Dixit, A. et al. Sequence and structure signatures of cancer mutation hotspots in protein kinases. PLoS One 2009;4(10):e7485. Feng, Y., Kloezekowski, A. and Jernig, R.L. Four-body contact potentials derived from protein datasets to discriminate native structures from decoys. Proteins 2007;68(1):35-66. Ferreiro, D.U. et al. Localizing frustration in native proteins and protein assemblies. Proc Natl Acad Sci U S A 2007;104(50):19819-19824. Fiser, A. et al. ModLoop: automated modeling of loops in protein structures. Bioinformatics 2003;19(18):2500-2501. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Small GTPases in Cancer. Nucleic Acids Res 2011;39(Database issue):D945-950. Gorfe, A.A., Grant, B.J. and McCammon, J.A. Mapping the nucleotide and isofrom-dependent structural and dynamical features of Ras proteins. Structure 2008;16(6):885-896. Grant, A.R. et al. Assessing the gene-disease association of 19 genes with the RSAsopaths using the ClinGen gene curation framework. Hum Mutat 2018;39(11):1485-1493. Grant, B.J., Gorfe, A.A. and McCammon, J.A. Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. PLoS Comput Biol 2009;5(3):e1000255. Hart, S.N. et al. Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. Genet Med 2019;21(1):71-80. Hobbs, G.A., Det, C.J. and Roosman, K.L. RS isoforms and mutations in cancer at a glance. J Cell Sci 2016;129(7):1287-1292. Hu, Z. et al. VIPDB, a genetic Variant Impact Predictor Database. Hum Mutat 2019;40(9):1202-1214. Hunter, J.C. et al. Biochemical and Structural Analysis of Common Cancer-Associated KRAS Mutations. Mol Cancer Res 2015;13(9):1325-1335. Ihaka, R. and Gentleman, R. A language for data analysis and graphics. J Computational and graphical statistics 1996;5(3):299-314. Ihle, N.T. et al. Effect of KRAS oncogene substitutions on protein behavior: implications for signaling and clinical outcome. J Natl Cancer Inst 2012;104(3):228-239. Ioannidis, N.M. et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 2016;99(4):877-885. Jenik, M. et al. Protein frustratometry: a tool to localize energetic frustration in protein molecules. Nucleic Acids Res 2012;40(Web Server issue):W348-351. Johnson, C.W. et al. The small GTPases K-Ras, N-Ras, and H-Ras have distinct biochemical properties determined by allosteric effects. J Biol Chem 2017;292(31):12981-12993. Karbassi, I. et al. A Standardized DNA Variant Scoring System for Pathogenicity Assessments in Mendelian Disorders. Hum Mutat 2016;37(1):127-134. Karchin, R. et al. Functional impact of missense variants in BRCA1 predicted by supervised learning. PLoS Comput Biol 2007;3(2):e26. Karczewski, K.J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of low-function-intolerance across human protein-coding genes. BioRxiv 2019;531210. Kiel, C. and Serrano, L. Structure-energy-based predictions and network modelling of RSAsopathy and cancer missense mutations. Mol Syst Biol 2014;10:727. Klee, E.W. and Zimmermann, M.T. Molecular modeling of LDLR aids interpretation of genomic variants. J Mol Med (Berl) 2019;57(4):533-540. Kocher, I.P. et al. The Biological Reference Repository (BioR): a rapid and flexible system for genomic annotation. Bioinformatics 2014;30(13):1920-1922. Kolde, R. and Kolde, M.R. Package ‘phetmap’. R Package 2015;1(7):790. Krijthe, J., van der Maaten, L. and Krijthe, M.J. Package ‘Rtsne’. In: GitHub. 2018. Kumar, R., Sukumar, S. and Barbacid, M. Activation of ras oncogenes preceding the onset of neoplasia. Science 1990;247(4959):1101-1104. Kumar, S., Clarin, D. and Gerstein, M. Localized structural frustration for evaluating the impact of sequence variants. Nucleic Acids Res 2016;44(21):10062-10073. Landrum, M.J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42(Database issue):D980-985. Lindsay, C.R. and Blackhall, F.H. Direct Ras G12C inhibitors: crossing the rubicon. Br J Cancer 2019;121(3):197-198. Liu, X., Tian, X. and Boerwinkle, E. daSNP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32(8):894-899. Ruocco, A. et al. daSNP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Hum Mutat 2015. Long, P.A. et al. De novo RRAGC mutation activates mTORC1 signaling in syndromic fetal dilated cardiomyopathy. Hum Genet 2016;138(8):909-917. Mihalov I., Res, I. and Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 2004;336(5):1265-1282. Milburn, M.V. et al. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science 1990;247(4959):939-945. Moom, K.R. et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2019;37(12):1482-1492. Munoz-Maldonado, C., Zimmer, Y. and Medova, M. A Comparative Analysis of Individual RAS Mutations in Cancer Biology. Front Oncol 2019;9:1088. Parra, R.G. et al. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrotaxis. Nucleic Acids Res 2016;44(W1):W356-360. Ponzone, L. and Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 2018;115(16):4164-4169. Porter, J.L. and Rose, G.D. A thermodynamic definition of protein domains. Proc Natl Acad Sci U S A 2012;109(24):9420-9425. Prior, I.A., Lewis, P.D. and Matton, C. A comprehensive survey of Ras mutations in cancer. Cancer Res 2012;72(10):2457-2467. Raunio, C. et al. The RASopathies. Annu Rev Genomics Hum Genet 2013;14:355-369. Rentzsch, P. et al. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47(D1):D886-D894. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405-424. Schymkowitz, J. et al. The FoldX web server: an online force field. Nucleic Acids Res 2005;33(Web Server issue):W382-388. Sceeb, P.H. et al. Biological properties of human c-Ha-ras1 genes mutated at codon 12. Gene 1984;312(5898):71-75. Simanshu, D.K., Nissley, D.V. and McCormick, F. RAS Proteins and Their Regulators in Human Disease. Cell 2017;170(1):17-33. Smith, M.J., Neel, B.G. and Ikura, M. NMR-based functional profiling of RSAsopathies and oncogenic RAS mutations. Proc Natl Acad Sci U S A 2013;10(12):4574-4579. Stenson, P.D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017;136(6):665-677. Tiddyman, W.E. and Rauen, K.A. The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. Curr Opin Genet Dev 2009;19(3):230-236. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. The Journal of Machine Learning Research 2014;15(1):3231-3245. Yang, Y. et al. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids 2013;44(3):847-855. Zimmermann, M.T. et al. Scriptural origins of misfolding propensity in the platelet adhesive von Willebrand factor A1 domain. Bioinformatics 2015;10(9):398-406. Zimmermann, M.T. et al. Assessing Human Genetic Variations in Glucose Transporter SLC2A10 and Their Role in Altering Structural and Functional Properties. Frontiers in Genetics 2018;9(576).