Bioinformatics, 37(10), 2021, 1367–1375
doi: 10.1093/bioinformatics/btaa972
Advance Access Publication Date: 7 December 2020
Original Paper

Structural bioinformatics

Structural bioinformatics enhances mechanistic interpretation of genomic variation, demonstrated through the analyses of 935 distinct RAS family mutations

Swarnendu Tripathi[1,2], Nikita R. Dsouza[1,2], Raul Urrutia[2,3] and Michael T. Zimmermann [1,2,4,5,*]

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

*To whom correspondence should be addressed.

Associate Editor: Valencia Alfonso

Received on July 29, 2020; revised on October 4, 2020; editorial decision on October 31, 2020; accepted on November 11, 2020

Abstract

Motivation: Protein-coding genetic alterations are frequently observed in Clinical Genetics, but the high yield of variants of uncertain significance 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.

Availability and implementation: All genomic and 3D scores, and markdown for generating figures, are provided in our supplemental data.

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 described 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 genetic 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 proteins, 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; Ihle 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 (Ioannidis 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 (Fig. 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 parameterized, 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 genomics.

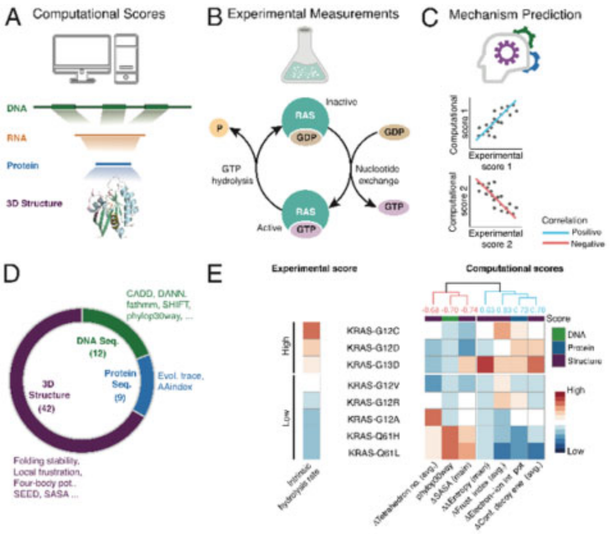

Fig. 1. Our process for integrating computational scores from multiple molecular levels with experimental data to interpret hotspot molecular mechanisms of genomic variants. (**A**) Multiple distinct molecules carry relevant information for directly interpreting the effects of genomic variants: the DNA 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 GTPase. (**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, $|R_{\text{Spearman}}|^3$ 0.6 (Supplementary 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_{Spearman}, respectively

2 Materials and methods

2.1. Selection of RAS family proteins and molecular modeling

We surveyed the PDB (Berman, 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 (Supplementary 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 Supplementary 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 (Supplementary 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 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 (Supplementary Table S1 and Supplementary Fig. 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 v5.0.0 (Kocher *et al.*, 2014) system for handling genomic data resources with custom scripts in the R programming language (Ihaka and Gentleman, 1996) for analysis. See Supplementary Materials for additional details. Because of codon degeneracy, multiple DNA changes could result in the same amino acid change. In these cases, protein scores are the same, but DNA sequence-based scores may differ from each other. Therefore, to be conservative, we chose the most severe score of the variants observed within each codon.

2.3. Protein variant scoring

For RAS missense variants, we combined scores at different molecular levels based on sequence (DNA and protein) and 3D protein structure using several available tools. At each level, we chose a subset of scores that each provide information about a different type of property (Supplementary Fig. S2A). These multiple molecular levels each carry information about the effect of a genomic variant but have not been used in combination. Because RAS is a GTPase, we note that these tools do not explicitly include ligand-protein interactions; we use them to focus on evaluation of the naturally occurring amino acids.

2.3.1. Variant sequence-based scoring

We selected 12 DNA sequence-based scores, including those that are the most frequently used in clinical genomics workflows, as well as some of the newest scores developed using data integration and machine learning that have the best overall predictive performance from large-scale assessments (Hart *et al.*, 2019; Ioannidis *et al.*, 2016; Liu *et al.*, 2015; Rentzsch *et al.*, 2019). See Supplementary Materials for additional details. We selected nine protein sequence-based scores by their breadth of physicochemical properties covered, low mutual correlations, and being commonly used in protein science, as described in depth in our Supplementary Materials.

2.3.2. Variant structure-based scoring

We selected 42 3D structure-based scores. Protein 3D properties have been studied in basic sciences for more than a century, but due to in part to their diversity of properties and low familiarity among geneticists, they have not been standardized and parameterized for use in clinical genomics workflows. Therefore, we gathered a diversity of 3D scores, including those that are highly established in basic science fields and recently developed novel 3D scores, to assess their similarity with the information available from DNA sequence-based scores. For example, we estimated the change in protein stability upon amino-acid substitution by calculating the change in folding free-energy, $_{\Delta\Delta}G_{\mathrm{fold}}$ using FoldX (Schymkowitz *et al.*, 2005) for each variant. We also computed 'local energetic frustration' for each RAS variant using the Ferreiro-Wolynes algorithm (Ferreiro *et al.*, 2007; Jenik *et al.*, 2012; Parra *et al.*, 2016) and estimated how favorable a specific contact (or an amino-acid residue) is relative to the set of all possible contacts (or an amino-acid residue) in that location compared to the energies of a set of 'decoy' states representing the molten globule configurations. To characterize the difference in regional folding cooperativity between the WT RAS and each variant, we employed a residue-specific implementation of structure-energy-equivalence-of-domains (SEED) algorithm (Porter and Rose, 2012). SEED parses proteins of known structure into their constituent thermodynamically cooperative components using residue-specific water $_{\rightarrow}$ 1 M urea-group transfer free energies (Zimmermann et al., 2015) to define thermodynamic subdomains of protein structures consistent with experimentally determined equilibrium folding intermediates. See Supplementary Materials for additional details.

All the 63 computational scores used in this study for the 935 missense variants from 7 RAS are included in the Supplementary Table S2.

2.4. Grouping variants by similarity across scores

Correlation among all 63 scores that we selected for consideration in our combined approach is shown in Supplementary Figure S2A. Using their mutual correlations, we selected 31 scores that were largely distinct representatives from the 63 (Supplementary Fig. S2B).

Analysis using t-distributed stochastic neighbor embedding (t-SNE) constructs 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 Section 2). 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 (Fig. 1A). We also assessed the available experimental data for RAS hotspot variants (Fig. 1B and Supplementary 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 (Fig. 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) (Fig. 1D and Supplementary Fig. 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 ($|R_{\text{Spearman}}| \leq 0.4$). 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 (Supplementary Fig. S2B). To guide our mechanistic characterization of the variants, we compared these 31 scores with direct experimental measurements, such as intrinsic GTP hydrolysis rates ($K_{\text{hydrolysis}}$) of the eight G12, G13 and Q61 hotspot variants (Fig. 1E). Noteworthy, we find that this method of experiment-guided parameterization of the computational scores for the hotspot variants (Fig. 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 (Supplementary 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 neoplastic 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. (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.

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 ($K_{\text{hydrolysis}}$)), GAP-stimulated $K_{\text{hydrolysis}}$, 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 (Supplementary Table S4). Next, we correlated these measurements with our multi-tier computational scores (Supplementary Fig. S3 and Supplementary Table S5). We identified seven computational scores (one DNA sequence, one protein sequence and five 3D-structure-based scores) that strongly correlate ($|R_{\text{Spearman}}| \geq 0.6$) with the intrinsic $K_{\text{hydrolysis}}$ values (Supplementary Table S5 and Supplementary Fig. S4A). Next, we represented these scores as a heatmap to better visualize their relationship with intrinsic $K_{\text{hydrolysis}}$ (Fig. 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 $K_{\text{hydrolysis}}$. 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 $K_{\text{hydrolysis}}$ by hotspot mutations ($P \leq 0.05$) (Supplementary 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 ($|R_{\text{Spearman}}| \geq 0.5$; Supplementary Fig. 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 $K_{\text{hydrolysis}}$ and nucleotide exchange rates are shown in Supplementary Figure 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 (Supplementary Fig. S4). We extended our mechanism investigation for all RAS variants using an unsupervised dimensionality reduction method, PHATE (see Section 2), 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 non-linear 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 versus non-hotspot variants. First, we focused our 2D PHATE analysis using the computational scores that strongly correlates with the experimental data of intrinsic $K_{\text{hydrolysis}}$ from the previous section (Fig. 1E and Supplementary Fig. S4C). In particular, we analyzed all the 935 variants from 7 RAS proteins and highlighted those 8 KRAS hotspot variants for which intrinsic $K_{\text{hydrolysis}}$ data is available (Fig. 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 (Fig. 2B). We note that in the 2D PHATE space G12, G13 and Q61 variants are distinct in Figure 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 (Fig. 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 Supplementary Fig. S4C) for all the 935 (from 7 RAS) and 493 (from HRAS, KRAS and NRAS) variants (Fig. 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 (Supplementary Fig. 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.

To explore whether any non-hotspot KRAS variants demonstrate similar effects in intrinsic $K_{\text{hydrolysis}}$ as compared with KRAS hotspot variants, we identified the 11 non-hotspot KRAS variants that are closest to the KRAS hotspot variants (N26Y and G138E near G12A, N26I near G13D and M1I, T35I, A59V, S65I, M72I, R123I, T124I and T127I near Q61L/H) in the 2D PHATE space shown in Figure 2B. These 10 amnio-acid residues are colored and projected based on their location in the *P*-loop, switch-I and switch-II regions in the 3D structure of KRAS along with the hotspot residues (Fig. 2E and F). Interestingly, these non-hotspot variants occur throughout the 3D structure indicating a need to functionally characterize non-hotspot alterations. In particular, the variants T35I is

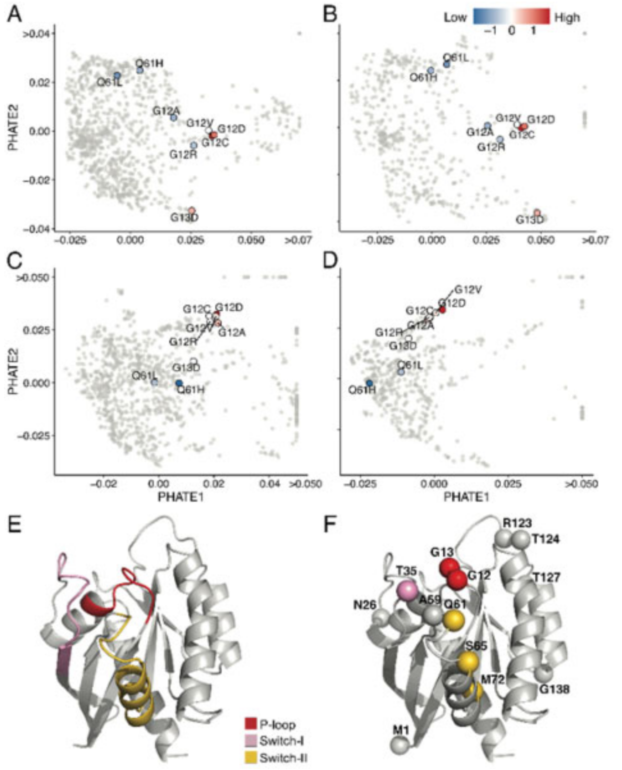

Fig. 2. KRAS non-hotspot variants computationally prioritized for effects on intrinsic and GAP-stimulated hydrolysis rates. We used the correlated computational scores (see Supplementary Table S4) to assess all non-hotspot variants for their potential to alter intrinsic and GAP-stimulated $K_{\mathrm{hydrolysis}}$, 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 $K_{\mathrm{hydrolysis}}$ of the KRAS hotspot variants (Fig. 1E). Eight hotspot somatic variants of KRAS are colored based on the intrinsic $K_{\mathrm{hydrolysis}}$ 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 $K_{\mathrm{hydrolysis}}$ (Supplementary Fig. S4C). In both (C) and (D), eight hotspot somatic variants of KRAS are colored based on the GAP-stimulated $K_{\mathrm{hydrolysis}}$ measurements relative to the WT. High (red) and low (blue) hydrolysis rates are indicated in the color bar. (**E**) 3D structure of KRAS (PDB: 4OBE) showing the sensitive regions, phosphate-binding loop (P-loop) (amino-acid 10-17), 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)

from switch-I, S65I and M72I are from switch-II, and A59V is at the nucleotide binding site (Fig. 2F). Other non-hotspot variants are just below switch-I (N26I) and on the far end of a loop that forms the nucleoside binding site and contains Noonan-associated germline mutations (R123I and T124I). From their pattern in computational scores that resembles hotspot variants, we believe these non-hotspot variants might have impact on intrinsic hydrolysis. Thus, these results emphasize the utility of more precise methods for interpreting the likely mechanistic effect of genomic variants.

3.3. Scores across multiple molecular levels identify which variants have similar effects

In the above sections, we focused on the computational scores that had the greatest correlations with quantitative experimental measures for hotspot mutations. Therefore, in this section, we expanded our analysis to consider patterns among 31 scores (Supplementary Fig. S2B), selected to maximally represent score diversity, and applied to all observed variants from seven RAS-family proteins. We performed data driven integrated analysis using an unsupervised non-linear dimensionality reduction process (t-SNE; see Section 2) and compared variants in a representative 2D space (Fig. 3A–C). The t-SNE-generated dimensions (t-SNE1 and t-SNE2 in Fig. 3A and B) are optimized in such a way that the scores of the variants, which are similar to one another in the raw high-dimensional data, are close in the 2D-reduced space.

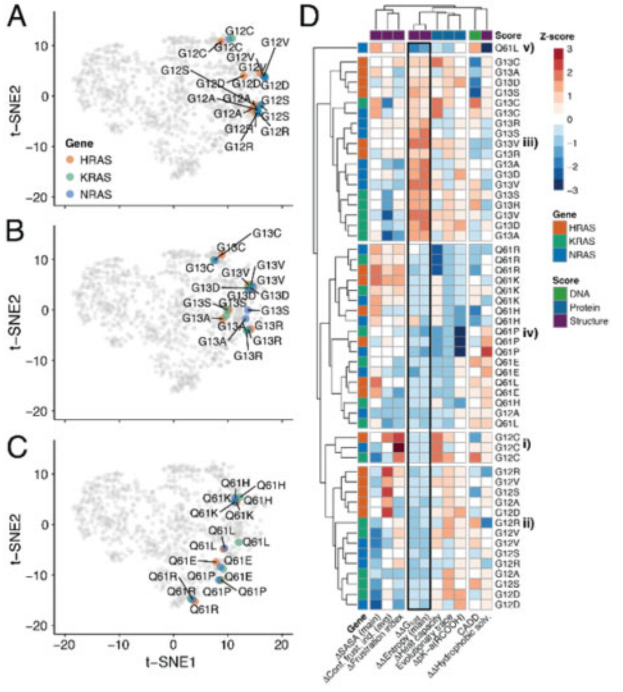

Fig. 3. All RAS variants assessed using our integrated scoring approach identify functional groupings among hotspot variants and demonstrate differences among non-hotspot variants. Having assessed global patterns by PHATE, we next characterized more nuanced local patterns among the variants using 2D 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^3 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 Supplementary Figure S6

In this section, first we focused on understanding the similarities and differences among the RAS hotspot variants and therefore t-SNE analysis is particularly useful as it emphasizes local relationships between the data points more than PHATE. We specifically analyzed the G12, G13 and Q61 missense gain-of-function variants in KRAS, HRAS and NRAS from within the large dataset. Notably, we found that all variants affecting the G12 and G13 hotspot residues are relatively close to one another in the t-SNE space, irrespective of the RAS member analyzed (Fig. 3A and B). However, we also find clear differences among hotspot variants. G12V 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 (Fig. 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 (Fig. 3A and B). We further noticed that Q61 variants (Fig. 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.

Next, we identified the most variable scores across genetic variants (median absolute deviation$_3$ 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 (Fig. 3D). We also represented all the 31 scores as a heatmap for the G12, G13 and Q61 variants in KRAS, HRAS and NRAS (Supplementary Fig. 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 ($\Delta\Delta G_{\mathrm{fold}}$). 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 Fig. 3D). Especially, the G13 variants are highly destabilizing (Median ± MAD; $\Delta\Delta G_{\mathrm{fold}} \sim 3.5 \pm 0.6$ kcal mol^{-1}), whereas G12 variants are either stabilizing or neutral ($\Delta\Delta G_{\mathrm{fold}} \sim -0.15 \pm 0.22$ kcal mol^{-1}). On the other hand, Q61 variants are moderately destabilizing ($\Delta\Delta G_{\mathrm{fold}} \sim 0.37 \pm 0.07$ 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 Fig. 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 '(iii)' in Fig. 3D] and show similar pattern in their scores, while G13C HRAS clusters with the other three HRAS G13 variants (G13A, G13D and G13S) (cluster 'iii' in Fig. 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 Fig. 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 Fig. 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 (Fig. 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 (Fig. 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 (Fig. 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 (Fig. 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.

Subsequently, we computed the correlations among all 63 computational scores from three molecular levels (Fig. 4D and Supplementary Fig. S2A). We found that protein structure-based scores have little-to-no correlation with the DNA sequence-based scores, overall (also see Supplementary Fig. 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

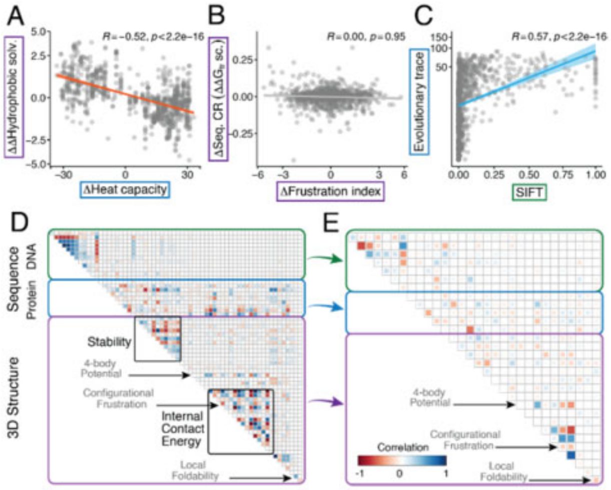

Fig. 4. Correlation among computational scores from multiple molecular levels for the 935 RAS variants demonstrates the distinct and underutilized value of 3D structure-based scores. (**A–C**) Spearman correlation (R_{Spearman}) among pair of computational scores indicating negative (A), neutral (B) and positive correlation (C). These three examples were chosen as exemplars for relationships between the scores from different molecular levels. (**D**) We used total 63 individual scores for the 935 protein variants to assess the differences among variants from 7 RAS proteins (KRAS, HRAS, NRAS, MRAS, RRAS, 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 Supplementary Figure 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 $|R_{\mathrm{Spearman}}|$. All 63 scores for 935 variants from 7 RAS genes are provided as Supporting Data (Supplementary Table S2)

the input experimental structure (Supplementary Fig. S8A) with mean R_{Spearman}=0.92 (Supplementary Fig. S8B). They also demonstrate high specificity (Supplementary Fig. 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 (Supplementary Fig. 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 (Fig. 4E and Supplementary Fig. 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 (Fig. 1E) and showed that mechanistic inferences can be made for why some variants alter enzymatic properties and others do not (Fig. 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 (Fig. 3). Next, we assessed a wide range of structure-based scores and applied them to all 935 RAS hotspot and non-hotspot variants (Fig. 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 $_{\Delta\Delta}G_{\text{fold}}$ associated with genomic variants (Cheng *et al.*, 2005; De Baets *et al.*, 2012; 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-family 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 (Supplementary Fig. 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 GTPase 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 (R01 DK52913 to R.U.), the Advancing a Healthier Wisconsin Endowment (to R.U.) and an endowment from the Linda T. and John A. Mellowes Foundation (to R.U.).

Conflict of Interest: none declared.

References

Andreoletti,G. *et al.* (2019) Reports from the fifth edition of CAGI: the critical assessment of genome interpretation. *Hum. Mutat.*, **40**, 1197–1201.

Angeles,A.K.J. *et al.* (2019) Phenotypic characterization of the novel, non-hotspot oncogenic KRAS mutants E31D and E63K. *Oncol. Lett.*, **18**, 420–432.

Bandaru,P. *et al.* (2017) Deconstruction of the Ras switching cycle through saturation mutagenesis. *Elife*, **6**, e27810.

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

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blackburn,P.R. *et al.* (2017) A novel Kleefstra syndrome-associated variant that affects the conserved TPLX motif within the Ankyrin repeat of EHMT1 leads to abnormal protein folding. *J. Biol. Chem.*, **292**, 3866–3876.

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

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

Cheng,J. *et al.* (2005) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinf.*, **62**, 1125–1132.

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

Clausen,R. *et al.* (2015) Mapping the conformation space of wildtype and mutant H-Ras with a memetic, cellular, and multiscale evolutionary algorithm. *PLoS Comput. Biol.*, **11**, e1004470.

Corominas,M. *et al.* (1991) ras activation in human tumors and in animal model systems. *Environ. Health Perspect.*, **93**, 19–25.

Cox,A.D. and Der,C.J. (2010) Ras history: the saga continues. *Small GTPases*, **1**, 2–27.

De Baets,G. *et al.* (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.*, **40**, D935–D939.

Dixit,A. and Verkhivker,G.M. (2014) Structure-functional prediction and analysis of cancer mutation effects in protein kinases. *Comput. Math. Methods Med.*, **2014**, 1–24.

Dixit,A. *et al.* (2009) Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS One*, **4**, e7485.

Feng,Y. *et al.* (2007) Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins*, **68**, 57–66.

Ferreiro,D.U. *et al.* (2007) Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA*, **104**, 19819–19824.

Fiser,A. *et al.* (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.

Fiser,A. and Sali,A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, **19**, 2500–2501.

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–950.

Gorfe,A.A. *et al.* (2008) Mapping the nucleotide and isoform-dependent structural and dynamical features of Ras proteins. *Structure*, **16**, 885–896.

Grant,A.R. *et al.* (2018) Assessing the gene-disease association of 19 genes with the RASopathies using the ClinGen gene curation framework. *Hum. Mutat.*, **39**, 1485–1493.

Grant,B.J. *et al.* (2009) Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS Comput. Biol.*, **5**, e1000325.

Hart,S.N. *et al.* (2019) Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genet. Med.*, **21**, 71–80.

Hobbs,G.A. *et al.* (2016) RAS isoforms and mutations in cancer at a glance. *J. Cell Sci.*, **129**, 1287–1292.

Hu,Z. *et al.* (2019) VIPdb, a genetic variant impact predictor database. *Hum. Mutat.*, **40**, 1202–1214.

Hunter,J.C. *et al.* (2015) Biochemical and structural analysis of common cancer-associated KRAS mutations. *Mol. Cancer Res.*, **13**, 1325–1335.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Ihle,N.T. *et al.* (2012) Effect of KRAS oncogene substitutions on protein behavior: implications for signaling and clinical outcome. *J. Natl. Cancer Inst.*, **104**, 228–239.

Ioannidis,N.M. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.*, **99**, 877–885.

Jenik,M. *et al.* (2012) Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res.*, **40**, W348–351.

Johnson,C.W. *et al.* (2017) The small GTPases K-Ras, N-Ras, and H-Ras have distinct biochemical properties determined by allosteric effects. *J. Biol. Chem.*, **292**, 12981–12993.

Karbassi,I. *et al.* (2016) A standardized DNA variant scoring system for pathogenicity assessments in Mendelian disorders. *Hum. Mutat.*, **37**, 127–134.

Karchin,R. *et al.* (2007) Functional impact of missense variants in BRCA1 predicted by supervised learning. *PLoS Comput. Biol.*, **3**, e26.

Karczewski,K.J. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

Kiel,C. and Serrano,L. (2014) Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol. Syst. Biol.*, **10**, 727.

Klee,E.W. and Zimmermann,M.T. (2019) Molecular modeling of LDLR aids interpretation of genomic variants. *J. Mol. Med. (Berl.)*, **97**, 533–540.

Kocher,J.P. *et al.* (2014) The Biological Reference Repository (BioR): a rapid and flexible system for genomics annotation. *Bioinformatics*, **30**, 1920–1922.

Kolde,R. and Kolde,M.R. (2015) Package 'pheatmap'. *R Package*, **1**, 790.

Krijthe,J. *et al.* (2015) Software Package 'Rtsne': T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. Version 0.15. https://github.com/jkrijthe/Rtsne.

Kumar,R. *et al.* (1990) Activation of ras oncogenes preceding the onset of neoplasia. *Science*, **248**, 1101–1104.

Kumar,S. *et al.* (2016) Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res.*, **44**, 10062–10073.

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–985.

Lindsay,C.R. and Blackhall,F.H. (2019) Direct Ras G12C inhibitors: crossing the rubicon. *Br. J. Cancer*, **121**, 197–198.

Liu,X. *et al.* (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.

Liu,X. *et al.* (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. *Hum Mutat.* **37**, 235–241.

Long,P.A. *et al.* (2016) De novo RRAGC mutation activates mTORC1 signaling in syndromic fetal dilated cardiomyopathy. *Hum. Genet.*, **135**, 909–917.

Mihalek,I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

Milburn,M.V. *et al.* (1990) Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. *Science*, **247**, 939–945.

Moon,K.R. *et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, **37**, 1482–1492.

Munoz-Maldonado,C. *et al.* (2019) A comparative analysis of individual RAS mutations in cancer biology. *Front. Oncol.*, **9**, 1088.

Parra,R.G. *et al.* (2016) Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.*, **44**, W356–60.

Ponzoni,L. and Bahar,I. (2018) Structural dynamics is a determinant of the functional significance of missense variants. *Proc. Natl. Acad. Sci. USA*, **115**, 4164–4169.

Porter,L.L. and Rose,G.D. (2012) A thermodynamic definition of protein domains. *Proc. Natl. Acad. Sci. USA*, **109**, 9420–9425.

Prior,I.A. *et al.* (2012) A comprehensive survey of Ras mutations in cancer. *Cancer Res.*, **72**, 2457–2467.

Rauen,K.A. (2013) The RASopathies. *Annu. Rev. Genomics Hum. Genet.*, **14**, 355–369.

Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.

Richards,S. *et al.*; on behalf of the ACMG Laboratory Quality Assurance Committee. (2015) 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.*, **17**, 405–424.

Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–388.

Seeburg,P.H. *et al.* (1984) Biological properties of human c-Ha-ras1 genes mutated at codon 12. *Nature*, **312**, 71–75.

Simanshu,D.K. *et al.* (2017) RAS proteins and their regulators in human disease. *Cell*, **170**, 17–33.

Smith,M.J. *et al.* (2013) NMR-based functional profiling of RASopathies and oncogenic RAS mutations. *Proc. Natl. Acad. Sci. USA*, **110**, 4574–4579.

Stenson,P.D. *et al.* (2017) 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.*, **136**, 665–677.

Tidyman,W.E. and Rauen,K.A. (2009) The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. *Curr. Opin. Genet. Dev.*, **19**, 230–236.

Van Der Maaten,L. (2014) Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, **15**, 3221–3245.

Yang,Y. *et al.* (2013) Structure-based prediction of the effects of a missense variant on protein stability. *Amino Acids*, **44**, 847–855.

Zimmermann,M.T. *et al.* (2015) Structural origins of misfolding propensity in the platelet adhesive von Willebrand factor A1 domain. *Biophys. J.*, **109**, 398–406.

Zimmermann,M.T. *et al.* (2018) Assessing human genetic variations in glucose transporter SLC2A10 and their role in altering structural and functional properties. *Front. Genet.*, **9**, 276.