Computational and Structural Biotechnology Journal 20 (2022) 2200–2211

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Structural bioinformatics enhances the interpretation of somatic mutations in KDM6A found in human cancers

Young-In Chi [a,b], Timothy J. Stodola [a], Thiago M. De Assuncao [a,b], Elise N. Leverence [a], Brian C. Smith [c], Brian F. Volkman [c], Angela J. Mathison [a,b], Gwen Lomberk [a,b,d], Michael T. Zimmermann [a,c,e,*], Raul Urrutia [a,b,c,e,*]

[a] Genomic Sciences and Precision Medicine Center (GSPMC), Medical College of Wisconsin, Milwaukee, WI, United States
[b] Division of Research, Department of Surgery, Medical College of Wisconsin, Milwaukee, WI, United States
[c] Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI, United States
[d] Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI, United States
[e] Clinical and Translational Sciences Institute, Medical College of Wisconsin, Milwaukee, WI, United States

ARTICLE INFO

Article history:
Received 22 February 2022
Received in revised form 18 April 2022
Accepted 18 April 2022
Available online 28 April 2022

Keywords:
Protein structure
Molecular dynamics
Genomic variation
Mutational impact analysis
KDM6A
Epigenetic regulator
Histone demethylase
Kabuki syndrome
Cancer

ABSTRACT

The histone demethylase KDM6A has recently elicited significant attention because its mutations are associated with a rare congenital disorder (Kabuki syndrome) and various types of human cancers. However, distinguishing KDM6A mutations that are deleterious to the enzyme and their underlying mechanisms of dysfunction remain to be fully understood. Here, we report the results from a multi-tiered approach evaluating the impact of 197 KDM6A somatic mutations using information derived from combining conventional genomics data with computational biophysics. This comprehensive approach incorporates multiple scores derived from alterations in protein sequence, structure, and molecular dynamics. Using this method, we classify the KDM6A mutations into 136 damaging variants (69.0%), 32 tolerated variants (16.2%), and 29 variants of uncertain significance (VUS, 14.7%), which is a significant improvement from the previous classification based on the conventional tools (over 40% VUS). We further classify the damaging variants into 15 structural variants (SV), 88 dynamic variants (DV), and 33 structural and dynamic variants (SDV). Comparison with variant scoring methods used in current clinical diagnosis guidelines demonstrates that our approach provides a more comprehensive evaluation of damaging potential and reveals mechanisms of dysfunction. Thus, these results should be taken into consideration for clinical assessment of the damaging potential of each mutation, as they provide hypotheses for experimental validation and critical information for the development of mutant-specific drugs to fight diseases caused by KDM6A dysfunctions.

1. Introduction

The *KDM6A* (Lysine-specific Demethylase 6A) gene encodes a histone H3K27 demethylase, also known as UTX, which forms part of the COMPASS complex and serves as a key epigenomic regulator involved in both normal and abnormal morphogenesis [1,2]. As such, alterations in this gene are found in Kabuki syndrome [3–5] and several tumor types, including multiple myeloma, bladder carcinoma, breast cancer, renal cell carcinoma, and pancreatic cancer [6–9]. Due to its key physiologic functions and disease-causal mutations, KDM6A is an important enzyme to investigate systematically, to determine the relative significance and underlying mechanisms for the wide range of observed genetic variation.

Abbreviations: 2OG, 2-oxoglutarate; COSMIC, Catalog of somatic mutations in cancer; dbSNP, Single nucleotide polymorphism database; DV, Dynamics variants; gnomAD, genome aggregation database; HAT, Hydrogen atom transfer; HMT, Histone methyltransferase; JmjC, Jumonji C domain; KDM6A, Histone lysine(K)-specific demethylase 6A; MD, Molecular dynamics; PDB, Protein data bank; Rg, Radius of gyration; RMSD, Root mean square deviation; RMSF, Root mean square fluctuation; SASA, Solvent-accessible surface area; SDV, Structural & dynamics variants; SNP, Single nucleotide polymorphism; SV, Structural variants; TCGA, The Cancer Genome Atlas; TPR, Tetratricopeptide repeat; VUS, Variant of uncertain (unknown) significance.

* Corresponding authors at: Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, 8701 W. Watertown Plank Rd., Wauwatosa, WI 53226, United States (M.T. Zimmermann). Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, 8701 W. Watertown Plank Rd., Wauwatosa, WI 53226, United States (R. Urrutia).

E-mail addresses: mtzimmermann@mcw.edu (M.T. Zimmermann), rurrutia@mcw.edu (R. Urrutia).

https://doi.org/10.1016/j.csbj.2022.04.028

Due to the medical relevance of KDM6A, several efforts have been directed toward the development of small molecules that inhibit the enzymatic demethylase domain of KDM6A [10–12]. These molecules can be chemically modified to fit into functional active or substrate binding sites. However, it remains unknown which mutations alter the demethylase domain and render KDM6A non-functional, defeating the purpose for which the therapy was developed. Correct annotations of these variations are vital for understanding its underlying molecular mechanisms and better diagnostic management and treatment for the patients. Importantly, only 15 KDM6A variants in the demethylase domain found to cause Kabuki syndrome have been carefully studied [13,14], while the complete repertoire of mutations found in cancer is broad and remains poorly understood. Consequently, this study was designed specifically to begin filling this knowledge gap in the field of genomics and precision oncology.

In the current study, we evaluated 197 somatic missense mutations (181 cancer-associated and 16 control variants) that affect 147 residues of the KDM6A catalytic domain, determining their effect on structural and dynamic properties. Critically, we reclassified each variant, based on the meta-scores derived from our multiparametric scoring system, for their effects on structure and dynamics, which underly their potential to not only alter enzyme function, but how drugs interact with the protein. Accounting for both structure and dynamics is critical because motions of the catalytic domain are required for function [15,16]. Our data indicate that KDM6A cancer variants can display functional disruptions in various ways. Some have subtle disturbance and alterations, but many others have severely damaging effects at the level of structure, dynamics, or both. The most damaging variants that affect multiple aspects of protein structure and dynamics are concentrated around the active site and substrate binding interface of the Jumonji C domain (JmjC) and the zinc-binding domain, reinforcing the known molecular mechanisms of KDM6A. These mutations disrupt the structural integrity, local geometry, chemical environment, and/or coordinated molecular motions. Together, the data resulting from these analyses combined with their quantitative evaluations provide new knowledge on how genomic variations impact KDM6A function. Thus, this information will have useful applications to score variants, to draw mechanistic inferences, and to help interpret results from drug development or testing.

2. Materials and methods

2.1. Selection of cancer-associated variants for the study

Across the genomic databases, there are over 300 somatic missense variants within the KDM6A catalytic domain among which we chose 197 variants on 147 residues. Selection of the cohort for this study was based on (i) inclusion of all variants listed in TCGA, (ii) highly expected 'damaging' and 'tolerated' variants from the initial structural analysis among additional entries in COSMIC, (iii) representatives from across the entire topology of the protein, and (iv) inclusion of all variants at or near the key functional sites.

In addition, a non-damaging control variant (H1060L) was selected because it was listed in ClinVar as a benign variant and as a single nucleotide polymorphism (SNP) in the general population database (dbSNP) with a high allele frequency (3.07×10^{-4}). Moreover, 13 most common gnomAD variants (with a high frequency greater than 1.5×10^{-5} in the general population) were added as potentially 'neutral' variants as they are expected to have no appreciable deleterious and pathogenic effects. On the other hand, two designed (non-natural) mutants, H1146A and E1148A on the two of the three Fe(II)-coordinating residues, were chosen as damaging controls because their loss-of-function activities are well documented in the literature [17–19]. Furthermore, twelve cancer somatic variants on 7 residues at the key functional sites served as additional damaging controls.

2.2. Preparation of the initial structure

High resolution (1.8 Å) crystal structure of a histone H3K27me3 peptide (17–33)-bound form of human KDM6A catalytic domains (PDB ID: 3AVR) was used in our study. This structure contains a Ni(II) ion (instead of the enzymatic Fe(II) ion) and the cofactor analog *N*-oxalylglycine at the active site as well as a structural Zn(II) in the zinc-binding domain. However, for our studies, the cofactor analog and the Ni(II) ion were replaced with the natural cofactor 2OG and the Fe(II) ion, respectively, to reconstruct the native-like active conformation. In addition, this structure is missing two loop regions (amino acids 902–910 and 1047–1078) due to high mobility, and they were filled in for our analysis using the Modeller program [20]. For missense variant analysis, substitutions were made within the Discovery Studio suite version 19.1 (Dassault Systèmes BIOVIA) by mutating the corresponding residue and selecting the side chain rotamer causing the least steric hindrance with the surrounding residues followed by energy minimization.

2.3. Protein folding energy and stability calculation

We assessed the stability of the mutated protein by the variant-induced changes in folding energy ($\Delta\Delta G_{fold}$) using FoldX [21] and the pH-dependent mutation energy protocol [22] implemented in the Discovery Studio suite (Dassault Systèmes BIOVIA). We used the energy minimized mutant structures for these calculations at pH 7.4 using the energy-minimized wild type structure (H3-unbound form) and introducing each substitution for calculation. After the preparation phase, the initial structures of the wild type and the generated mutants were subjected to a two-stage minimization process before being subjected to energy calculation. The predicted $\Delta\Delta G$ values, using both programs are in fairly good agreement (Supplementary Table S1 columns J & K).

2.4. Local structure perturbation measurement

We assessed the global and local structure perturbations by measuring the positional displacement of backbone atoms between the entire catalytic domains of wild type and mutant (global) and only the atoms near the residue of interest (local). For local structure perturbation, from the energy-minimized structures, any residues that reside within 10 Å radius from the mutation site were selected using PyMol (Molecular Graphics System, Schrödinger, LLC) and calculated for least-squared RMSD of the backbone atoms between the wild type and the mutant using Coot [23]. For global structure perturbation, entire backbone atoms were used for RMSD calculation between the structures.

2.5. pKa shift estimation

To investigate the possibility that some titratable residues may undergo protonation change upon single amino acid substitution of KDM6A, we performed the p*K*a calculations at pH 7.0 with DelPhiPKa [24] which is a surface-free Poisson-Boltzmann based approach to calculate the p*K*a values of protein ionizable residues. We first calculated the p*K*a values of titratable residues for the wild type and each mutant using energy-minimized structures. The p*K*a shifts then were calculated by subtracting the p*K*a values of the wild type and the mutant residues in both directions and summing up the differences. Loss or gain of titratable residues at the position

where cancer mutations occur and could dominate the cumulative p*K*a shift amount were not considered in calculations.

2.6. Molecular simulations

MD simulations were performed using the CHARMm36 all-atom-force-field [25] implemented in the Discovery Studio with a 2 fs time step. A simplified distance-dependent implicit solvent environment was used with a dielectric constant of 80 and a pH of 7.4, and no further parameterization of a non-standard residue (K27me3), cofactor, and metal centers. All MD simulations were carried out using periodic boundary conditions. Models were energy minimized for 5000 steps using steepest descent followed by 5,000 steps of conjugate gradient to relax the protein structure that was obtained under the stressed crystal environment. Each system of 10 replicates of wild type and each variant was independently heated to 300 K over 200 ps and equilibrated for 500 ps followed by 10 ns production simulation under NPT ensemble by changing the initial seed (100 ns total). Structures during unconstrained dynamics simulation were recorded every 10 ps to give a total of 1000 frames for analyses. This chemical timescale is enough for the side chain rearrangements in the protein's native state and to facilitate various conformations [26]. Total energy plots of the trajectories indicate that the systems can reach near equilibrium towards the end of the simulation. For final data analysis, one or two outliers (in some cases none) from each data set of 10 replicates that considerably deviate from the rest in RMSD plots and might represent the minor and rarer form of conformations (altogether 12% of the entire data) were excluded from averaging, and only the last 500 frames that have reached the near minimum total energy state were used. From 10 ns MD simulation, trajectory files were analyzed for structural impact by root mean squared deviation (RMSD), root mean square fluctuation (RMSF), and other measures such as time-dependent molecular interactions, radius of gyration (Rg), and solvent-accessible surface area (SASA). Trajectories were aligned to the initial wild type conformation prior to analysis. RMSD and RMSF values were calculated at the residue level for all atoms using the tools available within the Discovery Studio and the algorithms implemented in the Microsoft Excel program. Further analyses were carried out in the R programming language [27], leveraging the bio3d package [28]. Molecular visualizations were generated using PyMol (Molecular Graphics System, Schrödinger, LLC).

2.7. MD-associated parameter measurements

We first considered global changes in structure across our simulations and the evaluation of the structural drift was monitored by measuring all atom RMSDs from the initial structures as a function of time for all replicates and averaged (Supplementary Table S1). Next, additional information comes from the RMSFs of each amino acid, which highlights the flexible regions of the systems. Atomic RMSFs from the average structure in the trajectory were calculated for each residue and plotted as shown in our publication [13]. Finally, we computed the time-dependent binding free energy for all cases as well as radius of gyration (Rg) and dynamic solvent accessible-accessible surface area (SASA) using the protocols implemented in the Discovery Studio.

2.8. Identification of an additional effective MD-based metric at the active site

The lack of correlation of the active site interaction energies prompted us to seek out alternative MD-based metrics that might be more closely related to KDM6A catalysis. We focus more on the local chemical environment and geometry of the key reactive groups. The influence of the pK_a shift has been already noticed and presented in our publication [13]. However, the influence of local geometry has not been explored.

The key rate-determining step in the catalysis of the JmjC-containing enzymes is the hydrogen atom transfer (HAT) between the oxidized Fe(IV) of the reaction intermediate and the methyl group of the substrate [29] (Supplementary Fig. S1). Therefore, the Fe(II)-Me distance of the resting state plays an important role in the initiation and overcoming of the energy barrier for the reaction. To test its critical relevance to catalysis and potentiality as a protein-specific damaging effect metric within the active site, we monitored the distances between Fe(II) and the closest methyl group between which HAT must take place. Because there are three methyl groups in the lysine residue, we only measured the distance between the closest methyl carbon and the Fe(II) metal ion. Differences in these distance values between the wild type and the mutants show notable correlations with other scores (Supplementary Fig. S2), albeit lower than the other two MD-based scores (substrate/Zn interactions and RMSF), which prove that indeed HAT is a key molecular event that has been conserved throughout evolution. The lower correlation values than the other two might be due to inaccurate measurement of the distances stemming from the fact that the three equivalent methyl groups rotate/flip during MD simulation (Supplementary movie M1). The same behavior of the methyl groups was observed in the simulations of another KDM family member KDM7B [30] (Supplementary Fig. S2). This metric was included in the overall impact scoring based on these findings.

2.9. Time-dependent interaction energy calculation and HAT distance monitoring

Molecular interaction free energies were measured using the protocol implemented in Discovery Studio. This was done using the MD simulation trajectories and by selecting the protein and the interaction groups of interest. Non-bonded interactions were monitored and dynamic interaction energies (van der Waals and electrostatic energies) were calculated from using the CHARMm36 force field and the implicit distance-dependent dielectric solvent model. Fe(II)-Me distance monitoring was also done within Discovery Studio by selecting those atoms of interest. Because three methyl groups that rotate/flip during simulation (see the previous section and the Supplementary movie M1 and Fig. S2), we monitored all three and only took the shortest distance for averaging for all replicates.

2.10. Further use of the cross-correlation matrix to choose more appropriate scoring schemes

All MD-based data and the differences between the wild type and the mutants need to be properly scored for a more accurate impact assessment. While all measures of RMSD differences between the wild type and the mutants lack correlations, RMSF differences show notable correlations (Supplementary Fig. S3). Various measures of RMSF differences can be taken for potential scoring schemes and the best one needs to be determined. We used the cross-correlation matrix for this purpose and discover that either inversed Spearman or Pearson correlation coefficients between the RMSF plots of the wild type and the mutants (measures of non-native-like dynamics) show much better congruency than other measures such as average differences, absolute average differences, and residual differences (Supplementary Fig. S3, purple box). Thus, we chose the inversed Spearman correlation coefficients as the best scoring scheme for this metric for overall integration.

The cross-correlation matrix also helped to determine better scoring schemes for other metrics. For example, while decreases in folding/stability energy (destabilization) correlates well with other scores (thus, mono-directional), pK_a shifts in either direction (bi-directional) correlate much better with the other scores. Because the raw differences (both positive and negative differences) show poor congruency, we used absolute pK_a shift amounts as proper damaging scores [13]. Likewise, interestingly, substrate/Zn interaction energies show better correlations with other scores when the absolute differences are used (Supplementary Fig. S4). This indicates that the wild type exerts the optimal interactions with the substrate and any alterations (either loosening up or tightening the substrate binding) would disrupt the concerted dynamics during the catalysis (transition state formation) and the product release during the post-catalytic stage. On the other hand, the mutational effect of folding/stability and RMSD/RMSF seem to have only one-directional damaging effects. Similarly, we tested both possibilities for all other metrics and determine the proper scoring schemes that can be used for overall damaging impact assessment.

2.11. Overall impact classification of the variant

For overall impact scoring, we tentatively label 'benign', 'VUS' (variant of uncertain significance), or 'damaging' by referring to the control values and suggested threshold values for pre-classification (Supplementary Table S1). GnomAD variants with high allele frequency do not serve well as 'tolerated' or 'neutral' variants (see the main text). Using the suggested thresholds for sequence-based prediction tools as guidelines, we reclassified the variants based on meta-scores (0–0.2: tolerated, 0.2–0.3: uncertain, and 0.3–1.0: damaging). This results in a similar number of the tolerated variants with the sequence-based pre-classification (24 and 29 for pre-classification and reclassification, respectively). Moreover, all the damaging controls including the key functional disruptors can be classified as damaging variants, except one. Likewise, for 'molecular fitness' scores (without the sequence-based scores), the same threshold values were used, except that 'uncertain' ones were also regarded as damaging for a more complete cross-check with the pre-classified scores. Any variants that are predicted to be damaging at either category (structure or dynamics) or both would be ultimately assigned as damaging for 'molecular fitness' classification. More quantitative overall scoring schemes or a machine learning model will be considered when we have enough training sets from various proteins and supportive experimental data.

3. Results

3.1. Mutational landscape of KDM6A and its widespread genetic alterations in human cancers

Here, we describe the extended KDM6A mutational landscape across various human cancer types. First, we evaluated the KDM6A expression levels in cancer cells and generally they are lower or roughly equal to normal cells of these organs (Fig. 1A), suggesting that the dysfunction of this protein is not through aberrant expression but rather mutation-associated mechanistic alterations. For this study, we collected, organized, and pre-classified KDM6A variants by filtering somatic missense mutations reported in the public repositories, such as the Catalogue of Somatic Mutations in Cancer (COSMIC) [31] and the Cancer Genome Atlas (TCGA) [32], against the Genome Aggregation Database (gnomAD) [33] as a general population reference. KDM6A mutations are more common in bladder, lung, uterine, ovary, breast, and stomach tumors (Fig. 1B). Among the various mutations, we are focusing on the missense mutations as point mutations can be very instructive as to the functional/structural roles played by individual residues and provide critical information that can be utilized for selective therapeutic intervention. Indeed, over 60% of all somatic cancer variants contain missense or nonsense mutations (Fig. 1C).

In total, we report results derived from studying 197 variants on 147 residues within the catalytic domain (Fig. 2). These mutants consist of 181 cancer somatic variants, 3 control variants (H1060L benign, H1146A and E1148A damaging), and 13 neutral variants derived from the general population reference reported in gnomAD (Supplementary Tables S1–3). Distribution, frequency, and database sources (TCGA, COSMIC, and ClinVar entries) for all 181 cancer-related variants under study are shown in Fig. 2A and their locations mapped onto the KDM6A molecular structure are shown in Fig. 2B. Among these variants, twelve have been also reported in ClinVar [34], annotated as likely causes of Kabuki syndrome or 'not-specified' conditions (Fig. 2A). We did not find mutational 'hot-spot' regions affecting the sequence or the structure, regardless of cancer types (Fig. 2A-B). However, these cancer mutations in the KDM6A catalytic domain disrupt its histone demethylase activity, thereby leading to epigenomic alterations that contribute to cancer development [6–9]. Thus, to investigate their deleterious effects, we implemented a comprehensive sequence analysis (2D) combined with structural (3D) and dynamic-based (4D) approach to score the damaging impacts and lend insights into their mechanism of dysfunction.

3.2. Initial scoring of KDM6A variants using 2D sequence-based methods from clinical classification guidelines

Sequence-based annotation of pathogenicity is part of standard clinical practice, underlies information from genome variation databases, and is recommended by professional variant interpretation guidelines [35,36]. Thus, we initiated our studies in agreement with these guidelines, which we refer to as pre-classifications, using the widely available tools SNPs&Go [37], MutPred2 [38], PolyPhen2 [39], and Rhapsody [40], due to their proven performing rates in benchmark datasets [40–42]. PolyPhen2 combines sequence-based information with local structural features while Rhapsody incorporates structure-dependent properties and intrinsic dynamics derived from coarse-grained elastic network models. Although the results predicted by these algorithms are, for the most part, concordant, we find that there is a considerable degree of disagreement; over 40% of the variants (73 of 181 cancer variants and 79 of all 197 variants under study) show conflicting classifications (Fig. 2A and 2C). For example, SNPs&Go predicts that 49.7% of the cohort (98 of 197) are damaging while MutPred2 suggests that 79.2% of the cohort (156 of 197) are damaging, using their suggested threshold values (0.5 for both SNPs&Go and MutPred2). We also observe smaller, yet notable, discrepancies between the results obtained by PolyPhen2 and Rhapsody when their suggested threshold values are used (0.446 for PolyPhen2, and 0.5 for Rhapsody) (Supplementary Table S1). As shown in Fig. 2C, 47.7% (94 of 197 variants) are consensually predicted to be damaging and 12.1% (24 variants, not included in the diagram) are consensually predicted to be tolerated by these programs while 40.1% (79 variants) have conflicting predictions, thus considered to be variants of uncertain significance (VUS).

These conventional methods do not allow extensive evaluation of a mutated gene product in 3D nor account for the molecular dynamic environment (4D). Thus, we set out to study and establish a multi-parametric mechanistic-based assessment of the structural and dynamic features of the gene product with the goal of improving damaging predictions and gaining insights on molecular dysfunction at the atomic level (Supplementary Fig. S5).

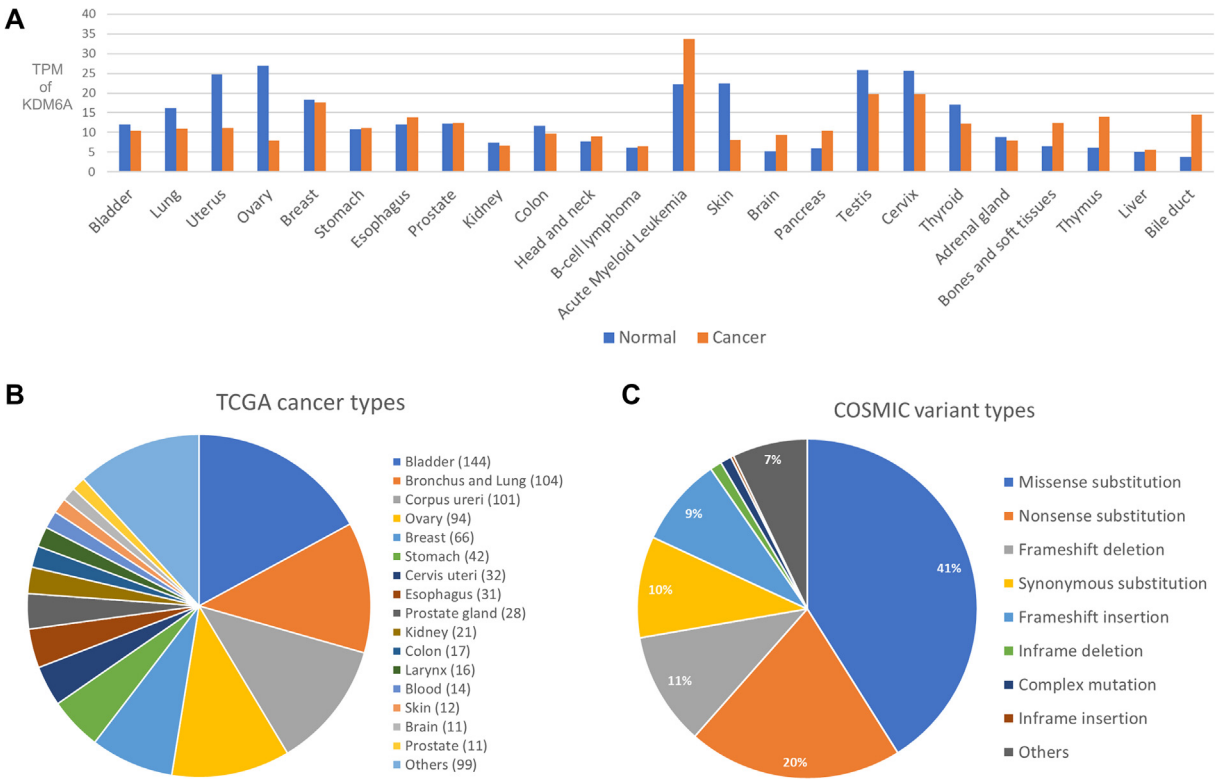

Fig. 1. *KDM6A* tissue-specific expression profile and cancer somatic mutation types. (A) *KDM6A* gene expression profile across all tumor samples and paired normal tissues. These data were extracted from the Genotype-Tissue Expression (GTEx) and Gene Expression Profiling Interactive Analysis (GEPIA) portals. TPM: Transcripts Per Million (B) Primary tissue types associated with KDM6A cancer mutations. A wide range of cancer types are observed with unequal prevalence. Although the figure was prepared with the TCGA data, similar distribution patterns of cancer types are observed in the COSMIC database. (C) Mutation types of KDM6A cancer somatic variants.

3.3. 3D structure-based scoring of KDM6A variants – defining structural variants (SVs)

To both complement and extend the results derived from sequence-based (2D) analytical tools, we leveraged the data derived from the high-resolution crystal structure for the KDM6A catalytic domain (PDB ID: 3AVR). This 3D genomics approach applies distinct analytical tools to evaluate structure-related features, both universal and protein-specific. In addition, we considered the contribution of both global and local features to compute damaging effects through the integration of variously scored biophysical properties. These features include key structural elements, such as altered protein folding, protein stability, local/global conformations, and pK_a values of the ionizable residues from the static structure because KDM6A is an oxidative enzyme whose catalytic activities are very sensitive to local pH and oxygen concentration [43,44]. Combined, these parameters offer insight into the parameters related to protein structure and dynamics (biophysical mechanisms) and other parameters related to its enzymatic function (biochemical mechanisms).

Within the catalytic domain, the key functional sites, such as the active site (Fig. 2D), and the substrate binding interface (Fig. 2E) including the zinc-ligations (Fig. 2F), can be identified and several cancer mutations are found within them. First, as an oxidative enzyme, KDM6A uses 2-oxoglutarate (2OG) and the metal iron as cofactors for catalysis. Nine key residues participate in the interactions with these cofactors and the methylated histone H3 K27 residue (H3K27me3) of the substrate, among which three residues (Y1135, K1137, and N1156) are mutated in cancer patients (Fig. 2D). Secondly, like any other enzyme, specific H3 substrate binding plays an essential role in the initiation and faithful execution of catalysis [29,45]. While many backbone atoms also participate in substrate recognition, four key residues are involved in sequence-specific side chain interactions (substrate specificity), among which one residue (E999) is mutated in cancer patients (Fig. 2E). Finally, the zinc-binding domain plays a critical role in selective substrate recognition as it goes through a substantial local conformational change when the histone substrate is bound [46]. Among the four Zn-coordinating residues, three (C1334, C1358, and C1361) are mutated in cancer patients (Fig. 2F). Collectively, these variants on the key functional residues served as additional damaging controls to evaluate the performance of the metrics that are being tested in the current study. No other critical functional elements or sites within the catalytic domain were recognized by conventional bioinformatics tools such as linear motif analysis or protein–protein interaction knowledge databases.

One of the main effects that a mutation can have on a protein is to alter its structural stability, causing local misfolding and a higher propensity for the variant to be targeted by degradation pathways [47]. Thus, we used FoldX [21] and Mutational Energy (Stability) calculation of the local protein structure [22] to estimate the energy differences between the wild type and each variant. As expected, the largest differences are observed in the well-folded regions of the structure (both the core and surface residues), such as the JmjC and the zinc-binding domain (Supplementary Fig. S6). Next, we measured structural perturbation (both global and local within a 10 Å radius), and pK_a shift amount (both global and locally around the active site) by amino acid substitution to assess impact. They all served well as effective measures of functional disruption. The comprehensive results for these measurements are provided in Supplementary Table S1 and graphically represented in Fig. 3A. Thus, through these approaches, we define a sub-class of genetic variants that damage enzyme function by changing properties of the protein structure, which we term structural variants (SVs).

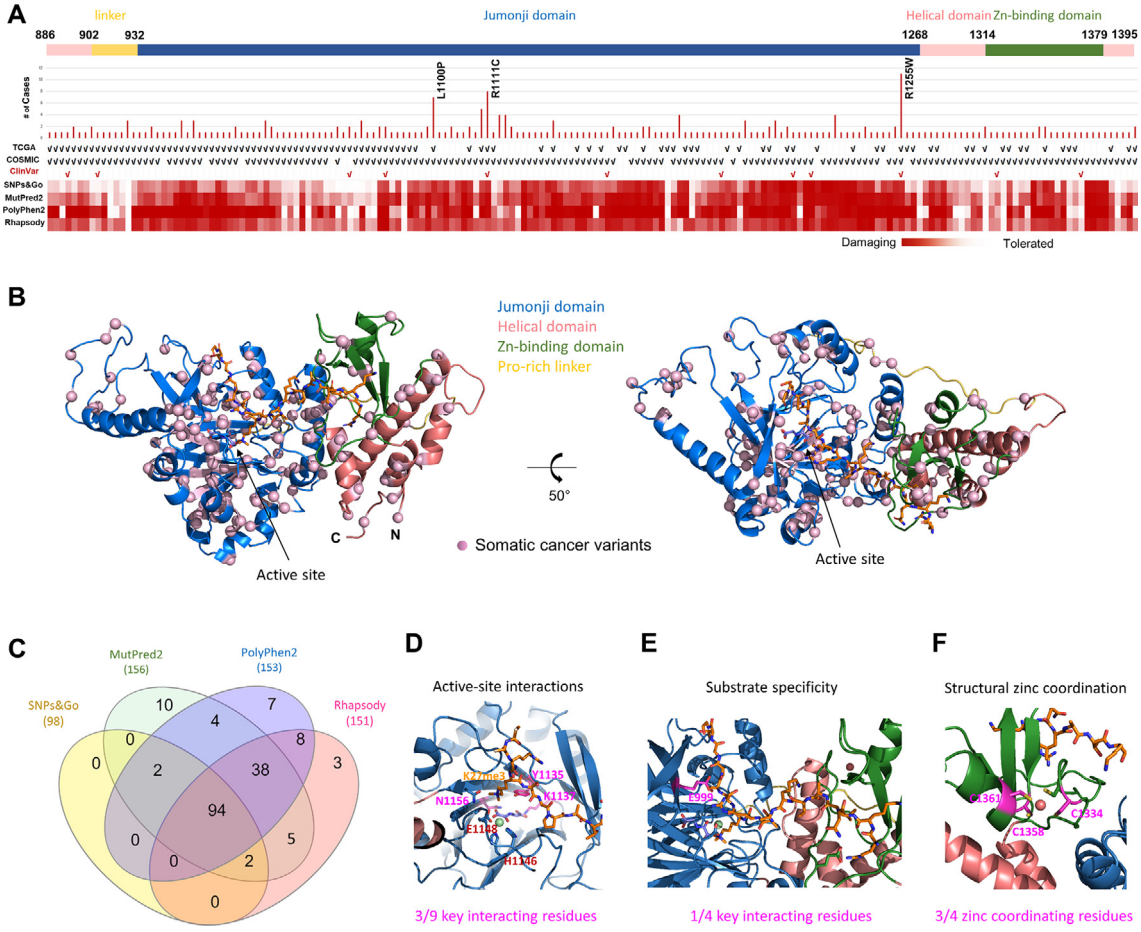

Fig. 2. Pre-classification of cancer-associated missense mutations and protein architecture reveal a diffuse landscape. (A) Sub-domain structure and distribution KDM6A missense mutations within the catalytic domain. The number of independent samples across TCGA, COSMIC, or ClinVar databases, harboring each missense mutation reveals that mutations spread out throughout the sequence, and the three mutations (L1100P, R1111C, and R1255W) have a high number of incidents. The impact predictions made by the genomics tools SNPs&Go, MutPred2, PolyPhen2, and Rhapsody are shown by small bars in the order of effect (damaging red to tolerated white). (B) Mapping of cancer-associated missense variants onto the KDM6A molecular structure. The catalytic domain is divided into the JmjC (blue) flanked by two additional sub-domains (helical domain: magenta and the zinc-binding domain: green) and a long flexible linker (yellow). The bound substrate is shown as ball-and-sticks while the catalytic domain is shown as ribbons. The color codes are identical to the ones used in Fig. 2A. (C) Venn diagram of the damaging variants predicted by each prediction tool, using the threshold value suggested by each program. Numbers of the damaging variants predicted by each program are indicated in parentheses. 47.7% variants (94 out of 197) share consensual damaging predictions while 40.1% (79 of 197) have conflicting predictions. The consensual tolerated variants (12.2%, 24 of 197) are not shown in this diagram. (D-F) Zoomed views of the key functional regions of KDM6A: the active site (D), the substrate binding interface (E), and the zinc ion binding site (F). Several key interaction residues (pink) are mutated in the cancer patients and they were used as additional damaging controls in the study. Within the active site, two non-natural damaging control residues (H1146 and E1148) are also labeled in red. H3 histone peptide residues (orange) are shown as sticks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. 4D dynamics-based scoring of KDM6A variants – defining dynamics variants (DVs)

We further complemented sequence- and structure-based analytical tools with the algorithms that can reveal the key molecular motions of distinct mutants, with the goal of mechanistic-based dysfunction predictions. For this purpose, we performed all-atom molecular dynamics (MD) simulations and probed the unique time-dependent molecular motions. We previously characterized the dynamic properties of the wild type KDM6A, revealed that the overall structure displays plasticity and a coordinated movement while maintaining the compact arrangement of the subdomains [13]. The conformational ensemble during the simulation reveals that the core JmjC containing the active site remains relatively still (less than 0.7 Å displacement), while the surrounding regions display more active motions (Supplementary movie M2). In addition to the mobile loops, the zinc-binding domain exhibits highly dynamic motions thought to be critical for substrate recognition and binding [13,46]. Furthermore, substrate binding is highly coordinated with the active site dynamics and synergistically influences the catalytic activities. In our simulations, metal ions and cofactors retain their bound state and ideal geometry with the ligating residues. Thus, there are protein-agnostic and protein-specific measures that we track over time to identify a set of mutations that we refer to as dynamics variants (DVs).

For MD-based analysis of the mutants, we measured (i) overall root-mean-square deviation (RMSD) of the conformations, (ii) root-mean-square fluctuation (RMSF) for individual residues, (iii) time-dependent interaction energies between the protein and the reactive groups at the active site or the substrate, (iv) dynamic distance calculations between the atoms involved in the rate-limiting catalysis, and (v) other parameters such as radius of gyration (Rg) and dynamic solvent-accessible surface area (SASA) (Supplementary Fig. S5). For example, we computed interaction energies (or energy changes due to mutations) between the protein and the reactive groups in the active site as well as the substrate. We found that grouping interaction energies into two key functional sites (active site cofactors and methyl group interactions, or substrate

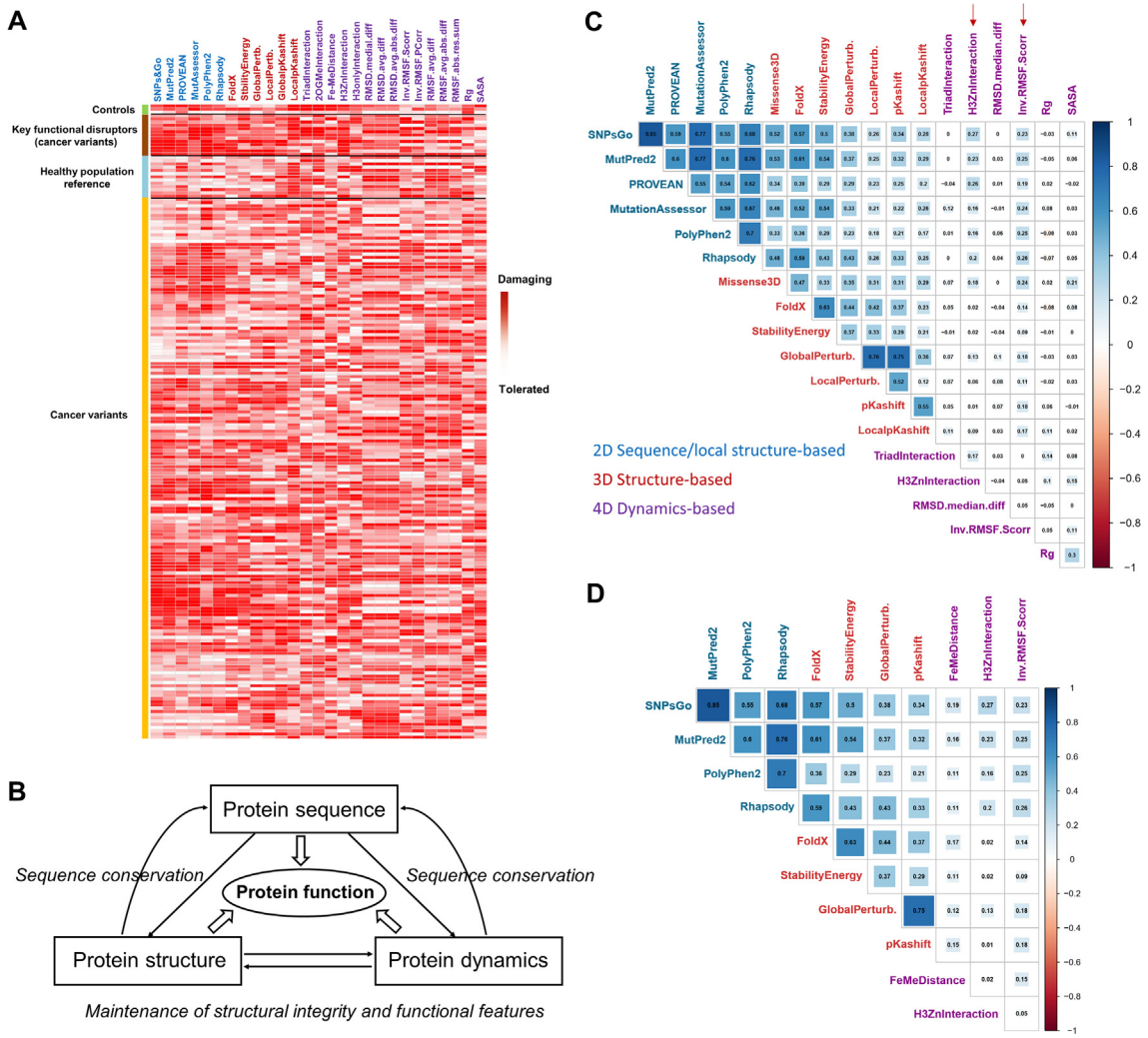

Fig. 3. Initial scoring and identification of congruent and more functionally relevant MD-based metrics by the cross-correlation matrix of the impact scores. (A) Heatmap of the initial raw scores by individual metrics. Because the original scores have different units and ranges in each column, they have been simply ranked to be equally scaled. Damaging scores are indicated by the intensity of red color. The controls are listed in the order of two well-known damaging and one benign mutant. The twelve key functional disruptors refer to mutations found right at the key functional residues. The contrast between the key functional disruptors and the gnomAD general population references are quite noticeable for sequence- and structure-based scores, but only for some selective MD-based scores that were identified by the subsequent congruency analysis. (B) Inter-relationship or dependence among protein sequence, structure, and dynamics for proper function. (C) Cross-correlation matrix of the scores from a comprehensive assessment. Among the MD-based scores, time-dependent substrate interaction-zinc ligation energy and RMSF (indicated by arrows) initially stand out to have notable congruencies with other sequence- and structure-based scores. (D) Cross-correlation matrix of the scores from the finally chosen metrics for meta-score calculations that are concordant and functionally relevant, thus have been evolutionally conserved. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and Zn interactions), instead of individual interactions, showed more consistent values among technical replicates and better correlations with other scores, presumably due to their coupled and coordinated nature. The largest differences in substrate/Zn interactions were detected in the zinc-binding domain (Supplementary Fig. S7), which is critical for substrate recognition and binding [46]. Additional minor changes were also observed around the active site, reinforcing the fact that substrate binding is highly coupled to the active site dynamics (Supplementary Fig. S7). The largest differences in molecular fluctuations were found near the active site although the active site has an overall rigid structure with peripheral regions have higher mobilities (Supplementary Fig. S8). The sensitivity of these residues to protein dynamics alterations heightens the importance of the residues peripheral to the active site to protein function. Among all MD-based metrics, more congruent and functionally relevant metrics were identified and used for overall impact scoring of each mutation on protein motions, leading to the classification of a subset of mutations as dynamics variants (DVs; Supplementary Table S3). This advance in 'molecular fitness' scores enables further classification of genomic variants with more mechanistic details.

3.5. Integration of individual scores using a multi-parametric meta-scoring approach

Although we used various metrics and measurements to gauge the damaging impact of genomic changes, we predict that not all structure- or dynamics-based metrics are directly related to molecular function. Thus, we used a cross-correlation matrix among the scores as a guidance to choose more effective and functionally relevant metrics for integration (Supplementary Text S1). This is based on the axiom that because protein sequence, structure, and dynamics are highly coupled for molecular function, those features affecting protein integrity and dynamics and critically needed for protein function more likely have undergone evolutionary conservation [48,49] (Fig. 3B). Indeed, our cross-correlation matrix of the

individual scores demonstrates there is a considerable degree of congruency among the measurements, indicating that collective interplay exists among all layers including dynamics to have led protein evolution and function (Fig. 3C). Structure-based scores show substantial congruency with sequence-based scores; however, not all MD-based scores display notable congruency with other scores. Strikingly, among the MD-based measurements, only substrate/Zn interaction energies and RMSF stand out as clearly congruent and functionally relevant metrics (see 'Materials and Methods' for the description of an additional effective MD-based metric identification and appropriate scoring schemes). RMSD and the active site interaction energies showed very little congruency, and their correlation values were close to zero (background noise appeared to be canceled out as more variants were added into the pool; see Supplementary Movie M3 for the progression of the matrix over time). When the data from all 197 variants were added into the pool, the correlation coefficients for substrate/Zn interaction energies and RMSF noticeably improved while those for the active site interaction energies, RMSD, Rg, and SASA came down to nearly zero, indicating that these numbers have converged to statistically significant values and no congruency exists for the latter four metrics.

These differences were quite striking and have several important implications. First, among the key interactions, only substrate/Zn interaction energies show congruency with other scores because, for JmjC-containing enzymes, only the substrate recognition dynamics serve as a rate-determining step [29,45]. Once the substrate is presented to the active site, catalysis readily takes place [50]. Moreover, the local conformations between the substrate-unbound and substrate-bound structures are nearly identical [46] and the catalytic mechanism of these proteins does not require additional regulatory elements near the active site. Thus, physical interactions within the active site might play a lesser role as the cofactors are firmly coordinated and the methylated lysine residue is readily presented to the active site once the substrate is bound. Instead, local chemical environment, such as pK_a and oxygen concentration, and local geometry, such as the distance between the key reactive groups, likely play more critical roles in catalysis [29] (see 'Materials and Methods' for the identification of Hydrogen Atom Transfer (HAT) distance as an additional effective MD-based metric). Although both global and local structure perturbations and pK_a shifts showed comparable congruency, only the ones with higher correlations (global measures) were chosen for final integration (Fig. 3C-D). Secondly, we reason why only RMSF differences and not RMSD differences show congruency is that the dynamic fluctuations or the protein-specific pattern of concerted and coordinated motions, rather than static measures of conformational deviations, are more closely related to protein's molecular function and conserved throughout evolution [51,52]. These results demonstrate that anomalous fluctuations of KDM6A can lead to non-active states and dysfunction. Overall, these findings indicate that the dynamic properties of the protein and its related protein-specific metrics, if chosen correctly, can serve as reliable indicators of protein function and dysfunction by disease mutations.

Based on these findings, for final integrated overall scoring, we chose four 2D sequence-based (SNPs&Go, MutPred2, PolyPhen2, and Rhapsody), four 3D structure-based (FoldX, stability, global structural perturbation, and overall pK_a shift), and three 4D dynamics-based metrics (molecular dynamics fluctuations, substrate/zinc-binding interactions, and HAT distances) that show notable congruency with sequence-based scores as shown in Fig. 3D. We limited the number of metrics to four to ensure that balanced contribution from each protein layer is given to the overall assessment although we found only three congruent metrics from the dynamics layer. For meta-scoring, because the impacting measurements for individual metrics are typically given in different units and have different ranges, we used Z-scores of the individual calculations and convert them into a zero to one scale that is commonly used by many sequence-based tools [53] before combining them (all eleven sequence-, structure-, and dynamics-based scores) with equal weights and averaging them for the final scores (Supplementary Table S2). These final meta-scores were used for the reclassification of the variants.

3.6. Reclassification of KDM6A genomic variants

Using the suggested thresholds for each prediction tool as guidelines, we reclassified the variants (0–0.2: tolerated, 0.2–0.3: uncertain, and 0.3–1.0: damaging, Supplementary Tables S2-3). The use of these criteria resulted in a similar number of the tolerated variants with the sequence-based pre-classification (24 and 32 for pre-classification and reclassification, respectively). The well-known damaging controls, H1146A and E1148A had meta-scores of 0.426 and 0.423, respectively while the benign control, H1060L had a meta-score of 0.095. Moreover, the key functional disruptors were all classified as damaging variants (meta-scores ranging from 0.395 to 0.650), except E999D (meta-score 0.295, thus still uncertain), perhaps due to the relatively conservative nature of this substitution. When these new threshold values were used, 69.0% (136 of 197) of variants belong to the damaging group and 16.2% (32 of 197) to the tolerated group while 14.7% (29 of 197) remain as VUS (Supplementary Table S2). This is a significant improvement from 40.1% VUS (79 out of 197) at pre-classification. Interestingly, however, relatively common variants from gnomAD with allele frequency greater than 1.5×10^{-5} in the general population [54] do not re-classify as 'tolerated' or 'neutral' variants wtih 5 out of 13 general population controls predicted to be damaging by our analyses (Supplementary Table S2). Perhaps, a higher allele frequency should be considered for any variants to represent the normal 'tolerated' or 'neutral' variants.

The reclassified variants were mapped onto the KDM6A molecular structure (Fig. 4A-B) and the aligned sequence of the KDM6 family members (Supplementary Fig. S9). While predicted damaging variants are concentrated near the active site within the JmjC and the substrate binding interface (Fig. 4A), the predicted tolerated or uncertain variants are mostly found on the fringe of the catalytic domain, away from the active site (Fig. 4B). However, we also discovered that the variants located immediately distant away from the active site and the substrate binding interface can impact protein function as they are known to serve as the 'second sphere', third, and beyond residues for enzymatic activities [29,55]. Coordinated functional motions through the physically neighboring residues are critical for this protein since multiple interactions at the active and substrate binding sites appear highly coupled. This inference is derived from the fact that disturbance by many mutations, including the damaging controls, had further rippling effects at the remote functional sites [13]. We found that the zinc-coordination disruptor C1334Y appeared to be the most deleterious one, having the highest meta-score of 0.65 as the zinc-binding domain plays a key role in the structural integrity and substrate binding. In addition, frequently observed variants from a wide range of cancer patients, such as L1100P, R1111C, and R1255W (Fig. 2A), also had relatively high meta-scores of 0.4613, 0.3531, and 0.5299, respectively. More importantly, when mapped onto the sequence alignment, 92% of all 102 damaging residues (136 variants) coincided with strictly conserved residues among the KDM6A family members (Supplementary Fig. S9). Only 8 damaging residues were found on the variable residues, and they can be explained by relatively drastic substitutions of the cancer variants (S893L, D980Y, K1053I, T1104P, V1205G, L1306S, D1340H, and D1382G). On the other hand, 25 tolerated residues (32 vari-

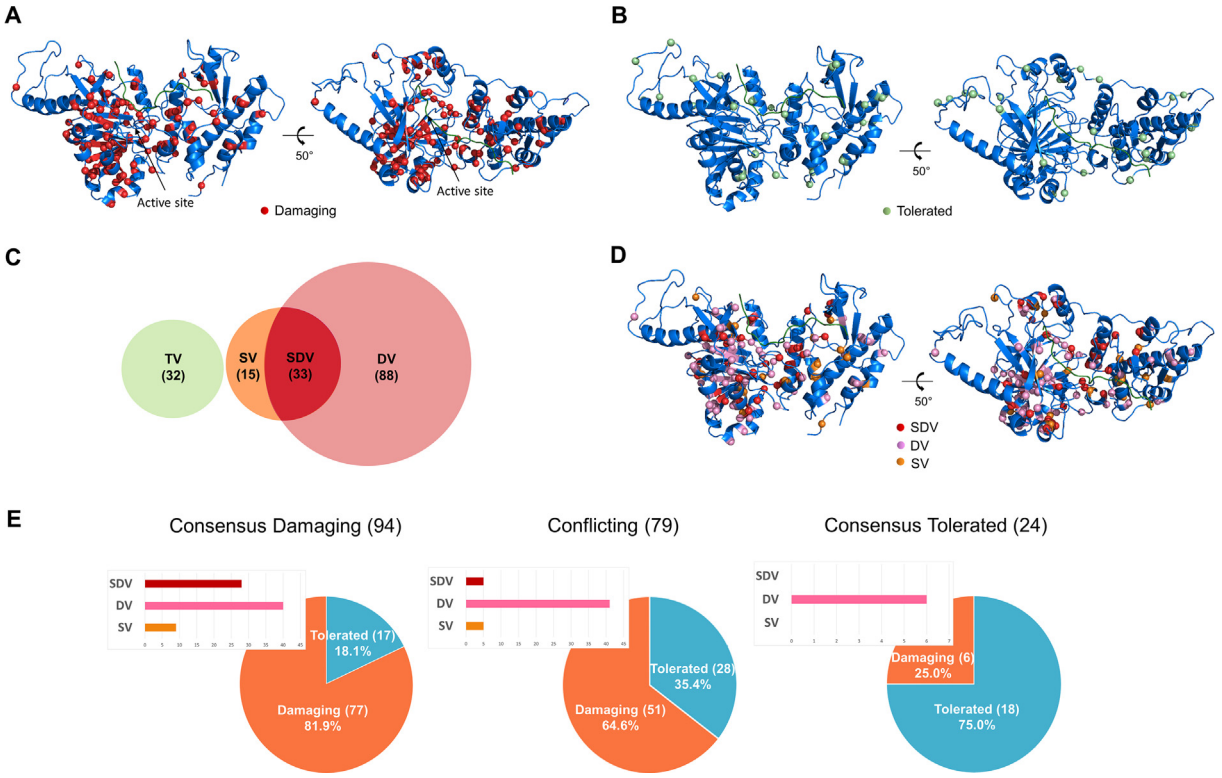

Fig. 4. Mapping of reclassified KDM6A cancer variants and comparison of conventional prediction tools to our comprehensive 'molecular fitness' assessment. (A-B) Damaging variants (A) and tolerated variants (B) based on meta-scoring reveal that the damaging variants (red) are concentrated near the active site (the jellyroll fold of the JmjC domain) and the substrate binding interface while the predicted tolerated variants (teal) are mostly found in the fringe of the catalytic domain. (C) Venn diagram of the tolerated (TV: green) and sub-grouped damaging variants, such as structural (SV: orange), dynamics (DV: pink), and structural & dynamics variants (SDV: red) based on our meta-scoring of all 197 variants. (D) Mapping of the sub-grouped damaging variants. The color codes are identical to the ones used in Fig. 4C. The most damaging variants (SDV) are all concentrated in the JmjC and the zinc-binding domains. (E) Comparison of conventional (sequence-based) prediction tools and comprehensive 'molecular fitness' (structural and dynamics-based) assessment for each pre-classified group. We compare the two classification results using a pie chart that indicates damaging versus tolerated for our new classification results, for each of the three pre-classification categories. The inset bar chart shows the balance between our three damaging categories. Overall, comprehensive assessments are in good agreement with the pre-classifications, but provide information of more specific mechanistic value. Confirmed damaging variants among the consensus damaging group by pre-classification (left chart) are altered in structure, dynamics, or both, while confirmed damaging variants among the consensus tolerated group (right) primarily affect protein dynamics. These types of mechanism-based interpretations should enable to resolve the conflicting variants (middle). Numbers of the variants in each group are indicated in parentheses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ants) mostly coincided with the variable residues, and only 5 tolerated variants (on 4 residues) were found on the strictly conserved residues (I1020M, S1071F, S1284L, and P1286S/L). Altogether, these findings support the overall effectiveness of our impact assessment scores.

3.7. Subgrouping of potentially damaging variants

Lastly, we also calculated 'molecular fitness' scores by considering only the structure- and dynamics-based scores (Supplementary Table S3), and benchmarked them against sequence-based scores (see the next section). Using individual 'molecular fitness' scores, we further classified the damaging variants into structural (SV), dynamics (DV), and structural & dynamics variants (SDV), as shown in Fig. 4C-D. The SVs are assigned using combined scores from protein folding/stability, structural perturbation, and pK_a shift while the DVs are assigned to those that represent substrate/Zn interactions, HAT distance monitoring, and RMSF scores. Altogether, our 'molecular fitness' evaluation shows that 48 variants disrupt at least one of the structural features and 118 variants disrupt at least one of the dynamic features (Fig. 4C and Supplementary Table S3). Among these, 33 variants (on 28 residues) disrupt both structural and dynamic properties of the protein. These 33 SDVs are predicted to be the most damaging variants, and among them, 28 (85%) are located within the JmjC and the remaining 5 (15%) are located within the zinc-binding domain (Fig. 4D), underscoring the key functional roles of these domains. Mapping of these most damaging variants onto the structure is shown in the Supplementary movie M4. No additional damaging sites such as remote allosteric sites were found. These findings highlight that KDM6A-specific conformational changes that are responsible for its function can be revealed by these measurements.

3.8. Comparison of the current biophysically enhanced approach with conventional sequence-based predictions

The ultimate question for the current study is how much additional prediction value we gain by including 3D structural- and 4D dynamics-based scores. Without actual wet bench experimental data, phenotypic information, or benchmarking data sets, we can tentatively evaluate how well the 'molecular fitness' scores match with the sequence-based scores by cross-checking them against each other. 'Molecular fitness' scores are calculated by integrating the structural and dynamics scores only (Supplementary Table S3). These two scores truly represent independent and unbiased measures as the 2D sequence-based scores represent mere predictions without offering mechanistic explanations for likely aberrant dysfunction. On the other hand, the 'molecular fitness' scores not only provide metric estimations but also yield interpretations or explanations underlying the mechanisms of variant dysfunction.

Our initial distribution of pre-classifications based on the sequence-based scores was depicted in Fig. 2D, in which 94 variants are consensually predicted to be damaging and 24 variants are consensually predicted to be tolerated. The remainder 79 (40.1% of all variants) have conflicting predictions, thus are regarded as VUS, representing the group where our approach has resolved the uncertainties and provided more accurate interpretations. The cross-checking results in each of the three pre-classification groups are shown in Fig. 4E. Overall, we observed a reasonable agreement of our 'molecular fitness' assessments with the pre-classifications, with some distinct differences. For instance, 18.1% and 25.0% of the damaging and tolerated consensus groups, respectively, displayed differences in predictions and structure/dynamics-based scores. This could be due to false predictions by the available sequence-based prediction tools or incomplete/incorrect interpretations by our current analyses. Mapping of the mismatched variants onto the structure did not hint at any salient causes. However, some pre-classified damaging variants, yet predicted to be tolerated by our 'molecular fitness' analysis such as Y1201H, A1203P, V1205G, R1213Q, G1215A, and D1216N are surface residues whose local properties or alterations thereof have not been thoroughly examined by our analyses. Furthermore, some other essential mutational impact assessments such as protein–protein interactions, protein expression, translocation, or post-translational modifications are not included in our current workflow. On the other hand, 'damaging' within the consensual pre-classified tolerated group could be due to insensitivity of sequence conservation enforced by functional dynamics. The mismatched variants turn out to be all dynamics disruptors (Fig. 4E third panel inset). For example, R922K and S925T mutations can be considered as conservative substitutions; however, our analyses confirm that their location in the proline-rich linker region can cause dynamics disturbance with loss-of-function effects [14] and thus, missed by sequence-based predictions. Therefore, while often regarded as tolerated, conservative substitutions can be more effectively scored if we consider their molecular dynamics. Overall, these data demonstrate the power of adopting an integrative approach over single amino acid conservation properties and these findings illustrate the potential benefits of our analyses although validation and further improvements are needed for more reliable predictions and interpretations.

4. Discussion

A preeminent challenge in human genetics remains the ability to accurately predict the molecular effects of genetic variants, as well as understand their impact on molecular and cellular functions. Current genomic guidelines for variant interpretation heavily rely upon co-observation, inheritance patterns, and evidence based on sequence and limited functional data [35]. However, the computational tools that lend supportive information to clinical classifications tend to produce high rates of false predictions [56], and their sensitivity and specificity suffer from the lack of mechanistic interpretations and insufficient data availability of their genotype-phenotype relationship [57]. Thus, there is an urgent need for improvement in computational approaches, and we aim to harness a more comprehensive and mechanistic-based computational assessment that is based on the overall 'molecular fitness' of the protein bearing the mutation.

In this study, we applied a mechanistic-based comprehensive approach that incorporates multiple aspects of protein structure, function, and dynamics of KDM6A. Nearly 200 cancer-associated missense variants within the catalytic domain, along with 16 selected controls, were chosen and characterized for structure- and dynamics-based impact predictions and interpretations. We analyzed each variant independently to offer a detailed picture of how curated missense variants may affect KDM6A enzymatic activities. Broader characterization of genomic variants with dynamic modeling, such as those identified in our cohort, has not been previously performed and represents a novel approach to understanding the functional effects of these changes. This level of detail is rarely considered in clinical genomics decision-making but is accessible to the computational technique of MD simulations in the lab and can lead to major strides within the field of medical genetics and genomic data science.

Our data indicate that KDM6A cancer variants display mechanistic disruption in various ways. Some have subtle disturbance and alterations, but many others have severely damaging effects at the structural or dynamics-level, or both. Overall, the pattern of functional evolution and functional disruption by these mutations reinforce the known molecular mechanisms of KDM6A. The most damaging ones that affect multiple aspects of protein structure and dynamics (33 SDVs of 197) are mostly concentrated around the active site and the substrate binding interface of the JmjC and the zinc-binding domain. These variants could be deleterious and contribute to tumorigenesis and progression as they can render the protein less functional and provide a selective growth advantage to cancer cells, although not all damaging variants are surely translated into pathogenic ones. On the other hand, most variants predicted to be tolerated are found distant from the key functional regions. These tolerated ones not observed in a generally healthy population may represent polymorphism or 'passenger' mutations in cancer patients with little or no effect in tumor progression. In other words, these mutations may occur as a consequence of tumorigenesis rather than a cause [58]. They could also represent polygenic mutants and accompany cooccurring cancer driver mutations in other related genes. Overall, cancer-associated mutants and their structure-dynamics-function relationships are well represented in the landscape of KDM6A genome variations.

The selected metrics and scoring schemes used in the current study produce scores that are in good agreement with the pre-classified scores and the expected values for various controls including the key functional disruptors. The cross-correlation matrix of the individual scores is very useful in identifying more effective/relevant metrics and scoring schemes. However, more reliable correlations of MD-based scores critically require enough samples (number of variants) and high-quality data (enough replicates and proper handling of the data). We did not notice any large differences in cross correlation values among the scores between the protein core residues and surface residues. From the static structure-based analysis, protein folding/stability and global/local structural perturbation are effective universal measures of damaging impacts. In addition, as a protein-specific measure, p*Ka* shift analysis is particularly useful for oxidative enzymes such as KDM6A. Likewise, from the dynamics-based analysis, time-dependent substrate/Zn interaction energy, HAT distance, and functional fluctuations are more closely tied to the molecular function as well as the sequence and structure conservation of KDM6A. Consequently, these measurements can serve as more reliable indicators of functional disruption. In the case of an enzyme, structurally coordinated dynamics enable the adaptation of the protein to binding substrates and to undergo allosteric transitions, while maintaining the native fold [16,59]. Thus, our findings reinforce the notion that fluctuations of the protein more strongly correlate with biological function (enzymatic properties) and have been evolutionally conserved.

Our findings also reaffirm the collective interplay among protein sequence, structure, and dynamics for protein function. Each layer exerts selective pressures on protein sequence that has maintained all aspects of protein structure, dynamics, and function.

More importantly, our data indicate that molecular dynamics-based evaluation is crucial to unveil the unique time-dependent information of mutations on functional mechanisms and dysfunction, which complement results obtained using existing tools. Thus, MD simulations, or other methods to calculate mutation-specific changes to protein motions, should become an integral part of genetic variation interpretation and meta-prediction approaches as recent years have seen its incorporation into advanced genomics analysis [60,61]. Due to their ability to simulate biochemical functions in time, unlike the 2D sequence-based scores, the specific qualities of the protein derived by our 3D and 4D genomics approach can be used in hypothesis driven testing. Although we have harnessed a limited set of MD-based metrics for the current studies, our results demonstrate that more effective and protein-specific metrics can be identified, and selected parameters can be used for each protein to probe protein-specific function. Additional standardized metrics such as protein surface features, protein network analysis, and principal component analysis can be added to expand the breadth of properties after testing their effectiveness on different proteins. Further optimization of MD simulation protocols, such as a better description of solvent environment and exploration of various production times and running parameters, will enhance the effectiveness of MD-based scores and the congruency with other scores.

Finally, it is likely that some genomic mutations will not affect the encoded 3D molecule but will instead be damaging due to altering transcription or translation. Further layers of information such as multi-tissue gene expression, protein abundances, post-translational modification, protein localization, etc., are not fully explored by the current approaches. Thus, studies, such as the current one, provide additional layers of information that need to be considered for enhancing our ability to interpret the effects of human genetic information. The inclusion of more effective metrics and scoring schemes and exploring multi-layered protein functions such as protein–protein interactions can improve the accuracy and sensitivity of the overall scores.

5. Conclusion

The current study significantly advances our understanding of precision oncology by providing insights into the damaging potential and mechanisms underlying the dysfunction of KDM6A mutations found in human tumors. This new knowledge will find application to diagnosis in precision oncology, mechanistic studies in cancer, and likely support a better understanding of epigenomic therapeutics. This type of analysis and nomenclature can be applied to other proteins and help better annotate the pathogenicity that can be curated into the public archives of human genetic variations for clinical applications. As more protein and protein complex structures become available, the widespread adoption of this approach will provide better diagnosis, risk assessment, and clinical guidelines for the observed mutants within the context of individualized medicine.

6. Data availability

MD Simulation data will be available upon request.

7. Competing interests

The authors declare that they have no competing interests.

8. Authors' contributions

RU devised the project, the main conceptual ideas, and the draft outline. RU, YC, and TJS designed the computational framework and analyzed the data. TMA, BCS, BFV, AJM, GL, and MTZ discussed the results, provided critical feedback, and helped shape the research and analysis. YC and RU were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgments

Funding information: This work was funded by the Theodore W. Batterman Family Foundation and the Advancing a Healthier Wisconsin Endowment to the Precision Medicine Simulation Unit of the Genomic Sciences and Precision Medicine Center at the Medical College of Wisconsin (to RU). This work was also in part supported by NIH grants R35GM128840 (to BCS) and R01DK052913 (to RU and GL).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.04.028.

References

[1] Swigut T, Wysocka J. H3K27 demethylases, at long last. Cell 2007;131(1):29–32.
[2] Gazova I, Lengeling A, Summers KM. Lysine demethylases KDM6A and UTY: The X and Y of histone demethylation. Mol Genet Metab 2019;127(1):31–44.
[3] Banka S, Lederer D, Benoit V, Jenkins E, Howard E, Bunstone S, et al. Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). Clin Genet 2015;87(3):252–8.
[4] Bogershausen N, Gatinois V, Riehmer V, Kayserili H, Becker J, Thoenes M, et al. Mutation update for Kabuki Syndrome genes KMT2D and KDM6A and further delineation of X-linked Kabuki syndrome subtype 2. Hum Mutat 2016;37(9):847–64.
[5] Miyake N, Koshimizu E, Okamoto N, Mizuno S, Ogata T, Nagai T, et al. MLL2 and KDM6A mutations in patients with Kabuki syndrome. Am J Med Genet A 2013;161A(9):2234–43.
[6] Tran N, Broun A, Ge K. Lysine demethylase KDM6A in differentiation, development, and cancer. Mol Cell Biol 2020;40(20).
[7] Wang L, Shilatifard A. UTX mutations in human cancer. Cancer Cell 2019;35(2):168–76.
[8] Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016;531(7592):47–52.
[9] van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, et al. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 2009;41(5):521–3.
[10] Zhang J, Ying Y, Li M, Wang M, Huang X, Jia M, et al. Targeted inhibition of KDM6 histone demethylases eradicates tumor-initiating cells via enhancer reprogramming in colorectal cancer. Theranostics 2020;10(22):10016–30.
[11] Heinemann B, Nielsen JM, Hudlebusch HR, Lees MJ, Larsen DV, Boesen T, et al. Inhibition of demethylases by GSK-J1/J4. Nature 2014;514(7520):E1–2.
[12] Andricovich J, Perkail S, Kai Y, Casasanta N, Peng W, Tzatsos A. Loss of KDM6A activates super-enhancers to induce gender-specific squamous-like pancreatic cancer and confers sensitivity to BET inhibitors. Cancer Cell 2018;33(3):512–26 e8.
[13] Chi YI, Stodola TJ, De Assuncao TM, Levrence EN, Tripathi S, Dsouza NR, et al. Molecular mechanics and dynamic simulations of well-known Kabuki syndrome-associated KDM6A variants reveal putative mechanisms of dysfunction. Orphanet J Rare Dis 2021;16(1):66.
[14] Petrizzelli F, Biagini T, Barbieri A, Parca L, Panzironi N, Castellana S, et al. Mechanisms of pathogenesis of missense mutations on the KDM6A-H3 interaction in type 2 Kabuki Syndrome. Comput Struct Biotechnol J 2020;18:2033–42.
[15] Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature 2007;450(7172):964–72.
[16] Vendruscolo M, Dobson CM. Structural biology. Dynamic visions of enzymatic reactions. Science 2006;313(5793):1586–7.

[17] Lee MG, Villa R, Trojer P, Norman J, Yan KP, Reinberg D, et al. Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. Science 2007;318(5849):447–50.
[18] Hong S, Cho YW, Yu LR, Yu H, Veenstra TD, Ge K. Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. Proc Natl Acad Sci U S A 2007;104(47):18439–44.
[19] Barrows D, Feng L, Carroll TS, Allis CD. Loss of UTX/KDM6A and the activation of FGFR3 converge to regulate differentiation gene-expression programs in bladder cancer. Proc Natl Acad Sci U S A 2020.
[20] Webb B, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol 2017;1654:39–54.
[21] Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320(2):369–87.
[22] Spassov VZ, Yan L. A pH-dependent computational approach to the effect of mutations on protein stability. J Comput Chem 2016;37(29):2573–87.
[23] Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 12 Pt 1):2126–32.
[24] Wang L, Zhang M, Alexov E. DelPhiPKa web server: predicting pKa of proteins, RNAs and DNAs. Bioinformatics 2016;32(4):614–5.
[25] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 2017;14(1):71–3.
[26] Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 2007;450(7171):913–6.
[27] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2020.
[28] Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22(21):2695–6.
[29] Ramanan R, Chatuvedi SS, Lehnert N, Schofield CJ, Karabencheva-Christova TG, Christov CZ. Catalysis by the JmjC histone demethylase KDM4A integrates substrate dynamics, correlated motions and molecular orbital control. Chem Sci 2020;11:9950–61.
[30] Chaturvedi SS, Ramanan R, Lehnert N, Schofield CJ, Karabencheva-Christova TG, Christov CZ. Catalysis by the non-heme iron(II) histone demethylase PHF8 involves iron center rearrangement and conformational modulation of substrate orientation. ACS Catal 2020;10(2):1195–209.
[31] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47(D1):D941–7.
[32] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375(12):1109–12.
[33] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581(7809):434–43.
[34] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46(D1):D1062–7.
[35] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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):405–24.
[36] Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017;19(1):4–23.
[37] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30(8):1237–44.
[38] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11(1):5918.
[39] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7(4):248–9.
[40] Ponzoni L, Penaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics 2020;36(10):3084–92.
[41] Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 2011;32(4):358–68.
[42] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.
[43] Gaweska H, Henderson Pozzi M, Schmidt DM, McCafferty DG, Fitzpatrick PF. Use of pH and kinetic isotope effects to establish chemistry as rate-limiting in oxidation of a peptide substrate by LSD1. Biochemistry 2009;48(23):5440–5.
[44] Chakraborty AA, Laukka T, Myllykoski M, Ringel AE, Booker MA, Tolstorukov MY, et al. Histone demethylase KDM6A directly senses oxygen to control chromatin and cell fate. Science 2019;363(6432):1217–22.
[45] Horton JR, Upadhyay AK, Qi HH, Zhang X, Shi Y, Cheng X. Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases. Nat Struct Mol Biol 2010;17(1):38–43.
[46] Sengoku T, Yokoyama S. Structural basis for histone H3 Lys 27 demethylation by UTX/KDM6A. Genes Dev 2011;25(21):2266–77.
[47] Scheller R, Stein A, Nielsen SV, Marin FI, Gerdes AM, Di Marco M, et al. Toward mechanistic models for genotype-phenotype correlations in phenylketonuria using protein stability calculations. Hum Mutat 2019;40(4):444–57.
[48] Liu Y, Bahar I. Sequence evolution correlates with structural dynamics. Mol Biol Evol 2012;29(9):2253–63.
[49] Konate MM, Plata G, Park J, Usmanova DR, Wang H, Vitkup D. Molecular function limits divergent protein evolution on planetary timescales. Elife 2019;8.
[50] Islam MS, Leissing TM, Chowdhury R, Hopkinson RJ, Schofield CJ. 2-Oxoglutarate-Dependent Oxygenases. Annu Rev Biochem 2018;87:585–620.
[51] Dong Z, Zhou H, Tao P. Combining protein sequence, structure, and dynamics: A novel approach for functional evolution analysis of PAS domain superfamily. Protein Sci 2018;27(2):421–30.
[52] Tiwari SP, Reuter N. Conservation of intrinsic dynamics in proteins-what have computational models taught us? Curr Opin Struct Biol 2018;50:75–81.
[53] Zeng Z, Bromberg Y. Predicting functional effects of synonymous variants: a systematic review and perspectives. Front Genet 2019;10:914.
[54] Itan Y, Casanova JL. Can the impact of human genetic variations be predicted? Proc Natl Acad Sci U S A 2015;112(37):11426–7.
[55] Hancock RL, Abboud MI, Smart TJ, Flashman E, Kawamura A, Schofield CJ, et al. Lysine-241 has a role in coupling 2OG turnover with substrate oxidation during KDM4-catalysed histone demethylation. ChemBioChem 2018;19(9):917–21.
[56] Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A 2015;112(37):E5189–98.
[57] Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal G, et al. Identification of cis-suppression of human disease mutations by comparative genomics. Nature 2015;524(7564):225–9.
[58] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446(7132):153–8.
[59] Ramanathan A, Agarwal PK. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol 2011;9(11):e1001193.
[60] Garg A, Pal D. Exploring the use of molecular dynamics in assessing protein variants for phenotypic alterations. Hum Mutat 2019.
[61] Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 2018;115(16):4164–9.