Predicting Oncogenic Missense Mutations

Xue Lei Boshen Wang Alan Perez-Rathke Wei Tian Chia-Yi Chou Yan-Yuan Tseng Jie Liang

***Abstract*—With the rapid progress of cancer genome studies, many missense mutations in populations of somatic cells of different cancer types and at different stages have been identified. However, it is challenging to understand the implications of these cancer-related variants. We have developed a computational method that integrates structural, topographical, and evolutionary information for assessments of biochemical effects and the extent of deleteriousness of the cancer-related variants. We have mapped somatic missense mutations from the Catalogue of Somatic Mutations In Cancer (COSMIC) to 3D structures in the Protein Data Bank (PDB). Our results show that a large portion of these missense mutations is located on protein surface pockets, which often serve as a structural and functional unit of cancer variants. We provide detailed analysis of several examples and assessment on the importance of these variants, including prediction of previously unreported cancer-variants, along with independent evidence from the literature. Furthermore, we show our predictions can inform on the functional roles and the mechanism of predicted cancer variants.**

***Index Terms*—cancer variants, somatic missense mutations, 3D structure, protein surface pocket.**

I. Introduction

A large amount of DNA mutations have been identified from genomic sequence of patients of various cancer types [1]. During cancer developments, many somatic mutations occur, some deleterious and some neutral. However, it was shown that people without cancer also have various mutations, while some cancer patients have no mutations [2]. Therefore, it is important to characterize these cancer-related variants and to distinguish which are deleterious and which are neutral.

There have been many efforts in addressing this issue. One approach is to identify deleterious variants based on unbiased statistical analysis of the mutation frequency. MutSig [3], OncoDriveClust [4], eDriver [5] are examples of efforts following this approach. These works have identified many deleterious cancer variants. However, due to the high frequency of background mutations and insufficient sample size, deleterious variants with low mutation frequency are difficult to distinguish from neutral variants. In order to identify rare variants, protein structure analysis has been employed widely in another class of approaches. These include HOTMAPS [6], HotSpot3D [7], e-Driver3D [8], 3DHotSpots [9]. These methods predict the deleterious mutations with low frequency. For example, the hypothesis of 3DHotSpots is that rare mutations which are closed to those known high-frequency variants are likely deleterious cancer-related mutations. Furthermore, additional rare mutations nearby are also tentatively to be deleterious mutations. Other approaches considered mutational processes [10], dynamic structural information [11], protein-protein interaction networks [12], protein family regions [5], etc. These results from different methods have overall little overlaps [5], since each method is using different information of the variants, suggesting they have strength on identifying different kinds of proteins and variants.

In this study we describe a novel approach that can aid in the assessment of the role of missense cancer variants, with the information of protein sequence conservation, and protein surface pockets incorporated, along with protein annotations such as functional domain, binding sites, mutagenesis, amino acid modifications, and post-translational modification sites. Below we first describe the procedure of mapping cancer variants to the 3D structures, we then discuss how protein surface pockets are computed and how their computations can help to evaluate missense variants for its deleteriousness that is cancer-related. We then give several detailed examples to illustrate how our method works, including the identification of well-known cancer-related variants. We also report the prediction of novel deleterious variants in the tumor suppressor PTEN, small GTPase RHOA and Serine/Threonine Kinase AKT1, that have not been reported in previous studies. We further point out evidence of our prediction from independent experimental literature. In addition, we show that our predictions can inform on the functional roles and the mechanism of cancer variants.

II. Methods

A. Cancer variants collection and mapping to 3D structures

We download all somatic missense cancer variants from COSMIC database in GRCh38, Feb. 2018, with a total of 469,544 cancer variants (https://cancer.sanger.ac.uk/cosmic). To map these cancer variants to a corresponding 3D structure, we use the HGNC (HUGO Gene Nomenclature Committee, https://www.genenames.org/) and SIFTS (Structure Integration with Function, Taxonomy and Sequence, https://www.ebi.ac.uk/pdbe/docs/sifts/) tables. For a given single amino acid variant, we first collect its gene name directly from the COSMIC database. We then map it to a single unique HGNC id, via a one-to-one mapping process. We then map the single unique HGNC id to a Swiss-Prot id, via the HGNC mapping table. Finally, we map a given Swiss-Prot id to a PDB structure, using SIFTS. At the same time, we map specific variants to PDB sequences via the SIFTS mapping table. Instead of locating all PDB structures, we seek to identify the best representative structures. Our goal is to identify possible key variants relevant to cancer among all the background mutations.

978-1-7281-0848-3/19/$31.00 ©2019 IEEE

B. Mapping the cancer variants to the protein surface pockets and assess the importance

We use the ComputedAtlas ofSurfaceTopography of proteins (CASTp) [13] to compute the protein surface pockets. Given a PDB structure, this algorithm returns the information of all surface pockets including the solvent-accessible surface area and volume, and the atoms that compose the corresponding surface pocket. We then map the cancer variants, that have been mapped to a PDB structure further to the surface pockets. We then obtain simple statistics for the given protein, including portion of variants that can be mapped to a 3D structure, and among these, the portion that can be mapped to a protein surface pocket.

For all the residues related to a structure, we assess their importance with structural attributes based on geometric and physicochemical properties. We developed a predictive model and used a predictor for assessing mutation effects. This predictor is trained using distinctive structural features of single amino acids, including its solvent-accessible area, polar and non-polar part, dihedral angles including phi and psi, conservation score from the entropy of sequence alignments among multiple species, as well as binary features such as whether it is a catalytic site and is a component of a salt-bridge pair. These computed properties can be further integrated in assessing mutation effects of a residue variant.

In our predictive model, the mutational effect $M(r_i) \in (0,1)$ of the variant at r_i is calculated as

$$M(r_i) = \alpha \left(\frac{1}{L_1} \sum\nolimits_i \frac{1}{L_2} \sum\nolimits_j s_{ij} \right) + \beta \sum_{m=1}^{n} \varphi_m \lambda_{im},$$

where α and β are weights for geometric and physicochemical terms, respectively. For variant at r_i, its geometric properties are expressed as $s_{ij} \equiv f_j c_j$, where f_j is the weighted alpha contact number [14], [15] to residue i computed from the alpha shape, c_j is the weight of residue computed from surface alignment, $L_1 = 6$ is the rotamer length, L_2 the number of residues with alpha contacts. The physicochemical properties are expressed as $\varphi_m \lambda_{im}$ where λ_{im} is a weight parameter, $\varphi_m \in (0,1)$ are scaled biophysico-chemical properties. Then $M(r_i) \geq \theta$, residue r_i is predicted to be deleterious in this weight scheme. We are able to identify the variants with the largest importance prediction score to be most relevant to cancer. We then assess the correctness of our method by examining annotations from UniProt and existing literatures. Below we give several detailed examples of our results.

III. Results

We examine several well-studied proteins known to have important roles in cancer development.

A. Tumor suppressor PTEN

PTEN is a tumor suppressor which acts as a dual-specific protein phosphatase. It dephosphorylates tyrosine-, serine- and threonine-phosphorylated proteins. It also acts as a lipid phosphatase. For this protein, we report here detailed analysis using the structure of PDB id 1D5R. This structure contains 324 residues out of the total protein length of 403 residues. Among all cancer variants that are related to this protein from COSMIC, 275 variants can be mapped into the protein surface pockets, 135 variants mapped outside the protein surface pockets, and 31 cannot be mapped to the structure (Fig. 1A). Furthermore, 15 residues are identified as driver mutation sites according to [9] (Fig. 1B). Among these, 7 residues are on the protein surface pockets and 8 residues are outside the surface pockets. All these 15 residues have a high score in our predictor, indicating high deleteriousness.

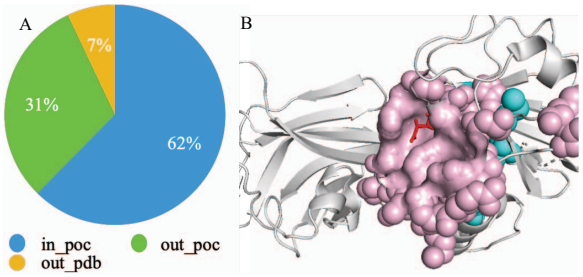

Fig. 1. Key residues identified in PTEN. (A) Statistics of cancer variants mapped to 1D5R. Blue, green and orange refers to the variants mapped into the surface pockets, outside surface pockets and outside this PDB structure, respectively. (B) Cyan spheres refer to the driver residues identified by [9]. Pink spheres refer to the additional residues we identified to be deleterious. Red refers to a ligand TLA.

There exist experimental evidence in the literature [16] for several key residues we identified as important but are not predicted in previous studies. Mutation D92A, which is not reported in [9], has a 700-fold reduction in phosphatase activity towards PtdIns(3,4,5)P3. Mutation H93A results in 75% reduction in phosphatase activity towards PtdIns(3,4,5)P3 and a modest reduction in phosphatase activity towards PtdIns(3,4)P2. Mutation K125M leads to reduced phosphatase activity towards PtdIns(3,4,5)P3, PtdIns(3,4)P2 and PtdIns(3)P. Mutation T167A leads to a 60% reduction in phosphatase activity towards PtdIns(3,4,5)P3. Both Q171A and Q171E have 75% reduction in phosphatase activity towards PtdIns(3,4,5)P3. All these functional changes may lead to the loss of protein phosphatase activity, leading to inability of PTEN protein to inhibit focal adhesion formation [16], which may lead to the loss-of-function as a tumor suppressor. These results indicate that our method can identify novel deleterious variants that are related to cancer.

B. Small GTPase protein RHOA

RhoA is a small GPTase that regulates the signal transduction pathway linking plasma membrane receptors to the assembly of focal adhesions and actin stress fibers. ROHA is involved in a microtubule-dependent signal, required for the formation of the myosin contractile ring during cell cycle cytokinesis and it plays an essential role in cleavage furrow formation. ROHA is also required for the apical junction formation of keratinocyte cell-cell adhesion. The whole protein has a length of 193. The 3D structure we choose, with the PDB

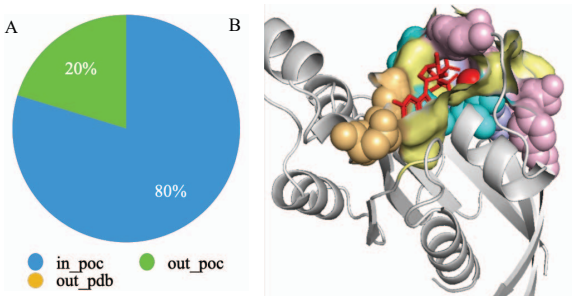

Fig. 2. Key residue identified in RHOA. (A) Statistics of cancer variants mapped to 1A2B. (B) Paleyellow shows the key surface pocket. Cyan, purple and orange spheres show the key residues we identified in the 3 GTP nucleotide binding region. Pink spheres show the identified key residues in effector region. Red refers to the ligand GSP and Mg+.

id as 1A2B, has one single chain and a total length of 182, covering the residue of the protein region from 1 to 181.

Results show that for all COSMIC cancer variants of RHOA, 59 variants can be mapped to protein surface pockets and 15 can be mapped outside pockets (Fig. 2A). All currently known cancer variants found on RHOA can be mapped to this structure. The key residues we predicted to be deleterious are shown in Fig. 2B. Analysis of the binding region and the effector region showed that residues G14, C16, G17 and T19 are in the first GTP nucleotide binding region, D59, T60, A61 and G62 are in the second GTP binding region, and K118 and D120 are in the third GTP binding region. Y34, T37, E40 and Y42 are in the effector region. All these residues are on the surface of the major surface pocket we identified and have high prediction score. Furthermore, except T37, E40, Y42, D59, T60 and A61, which have been reported in [9], the remaining 8 residues we predicted are novel.

Our results show that the surface pocket we identified serves as the key functional region. Among residues form this surface pocket, we have identified a number of deleterious residues, suggesting that collectively, they may have coordinated influence on cancer development.

C. Serine/Threonine Kinase AKT1

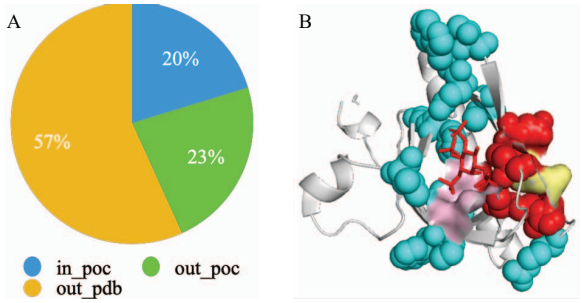

Fig. 3. Key residue identified in AKT1. (A) Statistics of cancer variants mapped to 1H10. (B) Paleyellow and lightpink show the key surface pockets. Cyan spheres show the predicted residues we identified. Red spheres refer to R15, E17 and E85. Red ligand refers to 4IP.

AKT1 is a serine/threonine-protein kinase that regulates many cellular processes including metabolism, and affect cell growth, proliferation, survival, and angiogenesis. Its structure (PDB 1H10) contains 125 residues. Among all cancer variants in COSMIC that can be mapped to this structure, 15 variants are mapped into its surface pockets, and 17 are outside the surface pockets (Fig. 3A). Among these mapped residues, 14 that are predicted to be highly deleterious, 6 of which are located on the surface pockets, and 8 are outside the surface pockets.

There are two connected surface pockets where many cancer variants are located (Fig. 3B). Residue R15 is predicted to be deleterious and participates in formation of both pockets. Predicted residues R15, E17 and E85, mapped inside the surface pocket (Fig. 3B, pink), are identified to be deleterious variants. Mutant R15Q is found in squamous_cell_carcinoma and E85K is found in malignant_melanoma, indicating the important roles of this surface pocket in skin cancers. Predicted residues W11, G33, F55, W80 and T81, mapped outside the surface pockets, are all found in adenocarcinoma cells. Among the 14 predicted residues, except for E17, L52, Q79 and W80, which have been identified in [17], the remaining 10 residues are newly discovered by our method.

The ligand, inositol-(1,3,4,5)-tetrakisphosphate (4IP, Fig. 3B, red), is a second messenger for Ca2+ modulation. Residues from 14 to 19 are all involved in 4IP binding [18]. Residues R15 and E17, predicted to be deleterious, are located in this region. Along with K20Q, variant E17K is known to cause loss of membrane localization [19]. Our prediction therefore identifies a functional unit formed by R15, E17 and E85, which are around these two important surface pockets identified for binding ligand 4IP and mediating Ca2+ transportation through the membrane. Thus, our predictions can inform on the functional roles and the mechanism of these cancer variants of this protein.

D. Computation of surface pocket for identifying key cancer variants.

Below we show quantitatively that the surface pocket information can help increase the success of our predictions. Fig. 4 shows the statistics of mutation sites that contains functional annotations. In Fig. 4A, we show the number of mutation sites compared to: (1) all mutation sites from the COSMIC database that can be mapped into the corresponding 3D structures; (2) mutation sites further restricted by the criteria of a prediction score larger than a certain threshold θ; and (3) mutation sites additionally restricted by the criterion of being mapped into a protein surface pocket. Since an oncogenic mutation causes either important loss-of-function or gain-of-function, it is reasonable to use annotations containing active sites, mutagenesis sites, modifications and functional regions as an indicator for the importance of the corresponding residue. As different proteins play different roles in cancer, we therefore use different threshold θ for assessing the prediction scores. The threshold is related to the mutation frequency and the length of the protein, which is correlated with the deleteriousness of the protein. In addition to PTEN, RHOA, and AKT1, we further include several other well-known cancer related

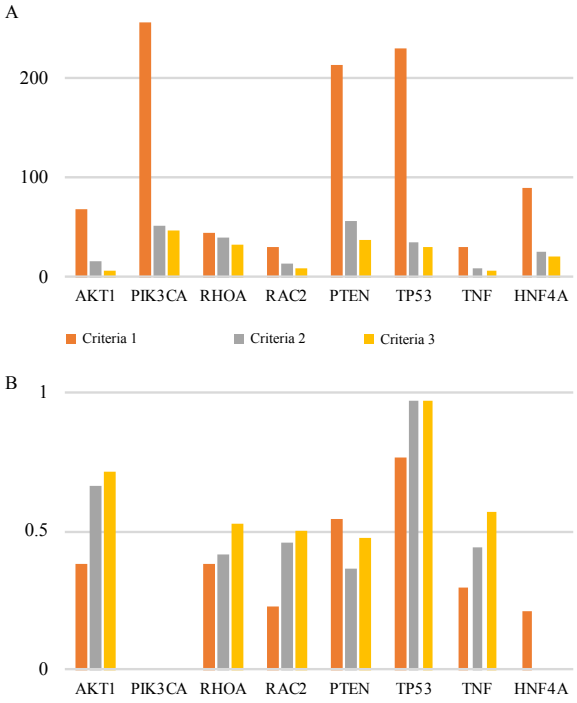

Fig. 4. Prediction and surface pocket information help on identifying key cancer variants. (A) Number of possible mutation sites. (B) Proportion of mutation sites with functional annotations. Parameter θ used for proteins in figure is AKT1: 0.7, PIK3CA: 0.85, RHOA: 0.7, RAC2: 0.9, PTEN: 0.95, TP53: 0.8, TNF: 0.6, HNF4A: 0.5.

proteins (*i.e.*, RAC2, PIK3CA, TP53, TNF, and HNF4A), details of which are dispensed with due to space limitation.

Fig. 4B shows the proportion of sites that contain at least one annotated important functional role. It is clear from Fig. 4B that with our criteria, the proportion that a site is of higher significance of cancer relevance generally increases when both the information of prediction score and the additional information of surface pockets are incorporated. Exceptions include PIK3CA and HNF4A, which contains very few mutation sites with a functional annotation.

IV. Conclusion

To summarize, we have developed a general method that can help to identify key residues relevant to cancer development from a large amount of background variants found in cancer patients. Our method integrates both sequence and structural information, and incorporates topographic information of protein surface pockets. We demonstrated that novel deleterious variants that have not been reported in previous studies can be identified using our method, which are supported by experimental literature. We have also shown that the computed surface pockets often form a structural and functional unit for deleterious variants. Furthermore, our predictions can inform on the functional roles and the mechanism of cancer variants. Our method is general, and can be applied to not only well-known proteins of oncogenes and tumor suppressor genes, but to other proteins as well.

Acknowledgment

This work is supported by NIH grants R01CA204962-01A1, R35GM127084 and R21AI126308.

References

[1] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, et al.,"Comprehensive Characterization of Cancer Driver Genes and Mutations", Cell, vol. 173, no. 2, pp. 371–385, 2018.

[2] R. Versteeg, "Tumours outside the mutation box", Nature, vol. 506, pp. 438–439, 2014.

[3] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, et al. "Discovery and saturation analysis of cancer genes across 21 tumour types", Nature, vol. 505, pp. 495–501, 2014.

[4] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes", Bioinformatics, vol. 29, pp. 2238–2244, 2013.

[5] E. Porta-Pardo, and A. Godzik, "e-Driver: a novel method to identify protein regions driving cancer", Bioinformatics, vol. 30, pp. 3109–3114, 2014.

[6] E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, and A. Godzik, "A pan-cancer catalogue of cancer driver protein interaction interfaces", PLoS Comput. Biol., vol. 11, pp. e1004518, 2015.

[7] B. Niu, A. D. Scott, S. Sengupta, M. H. Bailey, P. Batra, J. Ning, et al. "Protein-structure-guided discovery of functional mutations across 19 cancer types", Nat. Genet., vol. 48, no. 8, pp. 827–37, 2016.

[8] E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, and A. Godzik, "A pan-cancer catalogue of cancer driver protein interaction interfaces", PLoS Comput. Biol., vol. 11, pp. e1004518, 2015.

[9] J. Gao, M. T. Chang, H. C. Johnsen, S. P. Gao, B. E. Sylvester, S. O. Sumer, et al., "3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets", Genome Med., vol. 9, pp. 4, 2017.

[10] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, et al., "Signatures of mutational processes in human cancer", Nature, vol. 500, pp. 415–421, 2013.

[11] L. Ponzonia and I. Bahar, "Structural dynamics is a determinant of the functional significance of missense variants", PNAS, vol. 115, no. 16, pp. 4164–4169, 2018.

[12] H. Horn, M. S. Lawrence, C. R. Chouinard, Y. Shrestha, J. X. Hu, E. Worstell, et al., "NetSig: network-based discovery from cancer genomes", Nat. Methods, vol. 15, pp. 61–66, 2018.

[13] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "CASTp 3.0: computed atlas of surface topography of proteins", Nucleic Acids Res., vol. 46, pp. W363–W367, 2018.

[14] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar and S. Subramaniam, "Analytical Shape Computation of Macromolecules: I. Molecular Area and Volume Through Alpha Shape", Proteins Struct. Funct. Genet., vol. 33, pp. 1–17, 1998.

[15] X. Li, C. Hu, and J. Liang, "Simplicial edge representation of protein structures and alpha contact potential with confidence measure", Proteins, vol. 53, pp. 792-805, 2003.

[16] J. O. Lee, H. Yang, M. M. Georgescu, A. Di Cristofano, T. Maehama, Y. Shi, et al., "Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association", Cell, vol. 99, no. 3, pp. 323–34, 1999.

[17] A. Kamburov, M. S. Lawrence, P. Polaka, I. Leshchiner, K. Lage, T. R. Golub, et al., "Comprehensive assessment of cancer missense mutation clustering in protein structures", PNAS, pp. E5486–E5495, 2015.

[18] O. Dellis, S. G. Dedos, S. C. Tovey, Taufiq-Ur-Rahman, S. J. Dubel, C. W. Taylor, et al., "Ca2+ entry through plasma membrane IP3 receptors", Science, vol. 313, no. 5784, pp. 229–33, 2006.

[19] N. R. Sundaresan, V. B. Pillai, D. Wolfgeher, S. Samant, P. Vasudevan, V. Parekh, et al., "The deacetylase SIRT1 promotes membrane localization and activation of Akt and PDK1 during tumorigenesis and cardiac hypertrophy", Sci. Signal, vol. 4, no. 182, pp. ra46, 2011.