#5407 Use of computational algorithms to predict mutation effect in clinical setting

ESMO Congress 2022
9-13 September, 2022
Paris, France

Ekaterina Ignatova [1], Alexandra Lebedeva [2], Valentina Yakushina [2], Gregoriy Timokhin [2], Egor Veselovskiy [2], Vladislav Mileyko [2], Maxim Ivanov [2]

1. Research centre for medical genetics, Moscow, Russian Federation 2. Atlas Oncodiagnostics, LLC, Moscow, Russian Federation

Background

- Rare non-hotspot mutations in oncogenes are frequently identified during complex molecular profiling of tumors. Lack of functional and clinical data complicates interpretation of these variants as oncogenic or neutral.
- *In silico* algorithms may be useful for defining oncogenic status of previously uncharacterized mutations.

Results

Training dataset A total of 938 mutations in selected 42 oncogenes with consistent annotation across JAX and OncoKB databases were used as training dataset. 754 mutations were defined as *bona fide* oncogenic and 184 - as *bona fide* neutral mutations (Fig. 1A).

***In silico* neutral variants prediction** Different combinations of SIFT, PROVEAN, CADD, VEST4, CHASMplus, FATHMM, REVEL, MutPred, and MetaLR algorithms were combined to develop three sets of stringent criteria for high-confidence prediction of neutral status of mutations.

These allowed accurate prediction of 42 (23%) neutral mutation while 5 (2.7%) were erroneously predicted as neutral, as they were annotated as oncogenic in JAX/OncoKB (Fig. 1B). Three of them, EGFR p.I491M, p.G465R, p.S492R, located in extracellular receptor L-domain, are known to prevent binding of cetuximab resulting in therapy resistance, while no data about oncogenic effect of these variants were published.

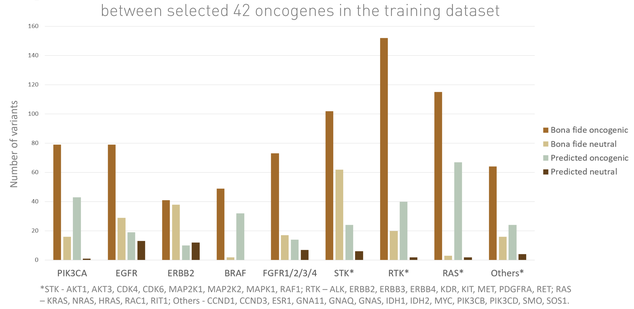
***In silico* oncogenic variants prediction** Additionally, 4 sets of criteria based on combinations of CHASMplus, VEST4, CADD, and PROVEAN algorithms were defined for prediction of oncogenic status which allowed correct prediction of 274 (36%) oncogenic mutations with 100% specificity (Fig. 1B).

Variant localization as a predictor of oncogenicity The variant proximity to the nearest hotspot and its location within a particular domain may be associated with its oncogenic status. In order to test this assumption variants the from training set were mapped to the protein structures from PDB database and Euclidian distances to the nearest hotpots obtained from the Cancer Hotspot database were calculated. Non-hotspot *bona fide* oncogenic and predicted oncogenic variants tend to localize in the spatial proximity to hotspots. Corresponding 3D-distribution curves with high accuracy resemble exponential distribution (R^2 = 0.8022 and R^2 = 0.9065, respectively), while neutral variant curve does not (R^2 = 0.3165) (Fig. 2A). Analysis of the distribution of mutations by domains showed that the majority of oncogenic variants are located in the protein tyrosine and serine/threonine kinase, as well as and Ras family domains. It is also noteworthy that a significant number of oncogenic mutations were located outside the known domains (Fig. 2B).

Retrospective analysis of clinical samples To evaluate the potential clinical applicability these sets of rules were applied to 130 unique mutations previously identified across 554 patients referred for complex tumor molecular profiling at our facility. This allowed for high-confidence prediction of neutral status of 20 mutations (100% are non-hotspot mutations) and oncogenic status of 44 mutations, including 5 non-hotspot mutations (ALK p.S1487L, RET p.W557I, KIT p.S746L, ERBB4 p.L428H and MET p.S637F mutations).

Observational dataset To further test the applicability of these rules in clinical setting sample from the MSK-IMPACT study containing mutations in 41 selected oncogenes were examined. This dataset included 5391 samples with a total of 9165 single nucleotide missense variant and 4121 unique variants.

Figure 1A Distribution of *bona fide* and predicted oncogenic and neutral variants between selected 42 oncogenes in the training dataset



*STK - AKT1, AKT3, CDK4, CDK6, MAP2K1, MAP2K2, MAPK1, RAF1; RTK – ALK, ERBB2, ERBB3, ERBB4, KDR, KIT, MET, PDGFRA, RET; RAS – KRAS, NRAS, HRAS, RAC1, RIT1; Others - CCND1, CCND3, ESR1, GNA11, GNAQ, GNAS, IDH1, IDH2, MYC, PIK3CB, PIK3CD, SMO, SOS1.

Figure 1B Confusion matrix for oncogenic and neutral variant predictions

Oncogenic variants

True positive 274	False negative 480
False positive 0	True negative 184

Neutral variants

True positive 42	False negative 137
False positive 5	True negative 749

Figure 2A 3D-distances of oncogenic, predicted oncogenic, and neutral variants from the nearest hotspot

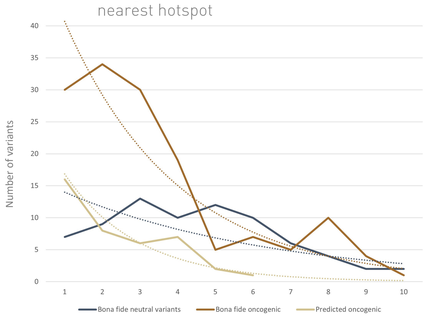


Figure 2B Distribution of predicted and oncogenic variants between domains

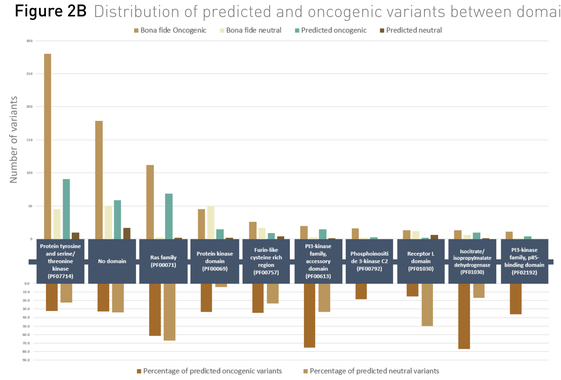
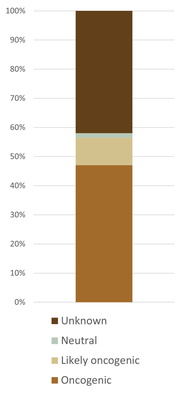


Figure 3

Annotations of variants in the observational dataset



Annotations and predictions of variants in the observational dataset

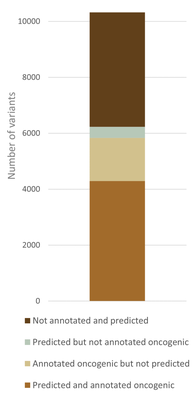
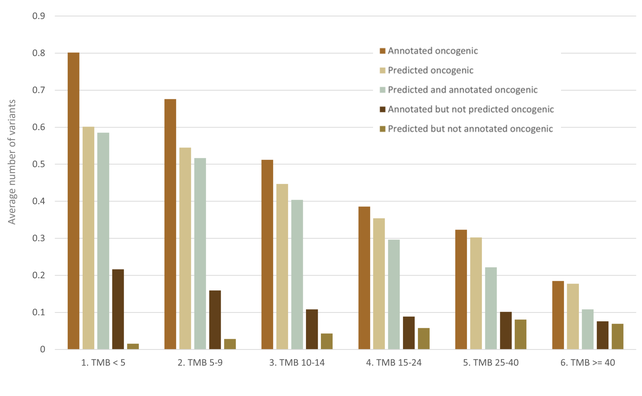


Figure 4

Distribution of an average number of oncogenic mutations in 41 oncogenes in samples by groups with different tumor mutational burden in the observational dataset



5361 variants were described as oncogenic or likely oncogenic in OncoKB knowledge base, 150 - as Neutral/Likely neutral/Inconclusive/Resistance, and 3654 had no interpretation.

Applying this set of rules allowed to predict 4731 (51.6%) variants as oncogenic. Amid them 4319 (91.3%) were described in OncoKB as oncogenic/likely oncogenic, 11 (0.2%) - as Neutral/Likely neutral/Inconclusive/Resistance, and 401 (8.5%) variants were not described in OncoKB (Fig. 3). Among patients included in MSK-IMPACT, 3.1% (n=166) had no known oncogenic single nucleotide variant in 41 oncogene but had at least one variant predicted to be oncogenic. Additionally, 6.5% of patients (n=350) had at least one predicted but not annotated oncogenic variant.

Variants were analyzed depending on TMB in corresponding sample. In the TMB-low subgroup (<5 mut/MB), the average number of known oncogenic mutations in 41 oncogenes per sample was high (0.8). This may indicate the increased likelihood of oncogenicity of the variants found with a low TMB when no other driver mutations are present (Fig. 4).

Methods

- JAX and OncoKB databases were used to obtain information on cancer mutations and create training set composed of *bona fide* oncogenic and neutral mutations.
- 13 *in silico* prediction algorithms were tested which resulted in the selection of CADD, CHASMplus, FATHMM, REVEL, MutPred, MetaLR, ProVean, SIFT, and VEST4 for the development of rules for the prediction of mutation effects.
- PDB and Cancer Hotspot databases were used to calculate distances between mutations and hotspots.
- Retrospective analysis of NGS data obtained from clinical samples of patients who have undergone comprehensive tumor molecular profiling at our facility.
- The list of oncogenes included in the analysis is listed in the Figure 1A.
- Data from MSK-IMPACT study was obtained from cBioPortal and used to compose observation dataset.

Conclusions

- Commonly used *in silico* algorithms can be combined for high-confidence prediction of mutation oncogenic or neutral status.
- Additional evidences such as spatial proximity to the nearest mutation hotspot, localization in the particular protein domain, tumor mutational burden in the sample, and presence of another driver events in the sample affect the likelihood that the variant is either oncogenic, or neutral.
- A combination of *in silico* algorithms, together with the set of listed evidences can be used to select patients for inclusion in a clinical trial or for off-label therapy in the absence of other suitable therapeutic options.

Corresponding author email address: maksim.v.ivanov@phystech.edu