TRACEvar: Prioritizing and interpreting pathogenic variants that underlie hereditary diseases by using machine learning and tissue contexts

Chanan M. Argov[1], Eric Sabag[1], Avigdor Mansbach[1], Yair Sepunaru[1], Emmi Filtzer[1], Yuval Yogev[2], Ohad Birk[2,3], Vered Chalifa-Caspi[4], Lior Rokach[5], Esti Yeger-Lotem[1,3]
[1] Department of Clinical Biochemistry and Pharmacology, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel. [2] Morris Kahn Laboratory of Human Genetics and the Genetics Institute at Soroka Medical Center, Faculty of Health Sciences, Ben Gurion University of the Negev, Beer Sheva 84105, Israel. [3] The National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel. [4] Ilse Katz Institute for Nanoscale Science & Technology, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel. [5] Department of Software & Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel.

Hereditary diseases affect millions of people around the world. Identifying disease-causing pathogenic variants out of thousands of candidate variants and understanding their mode-of-action is a major challenge and a crucial step in studying and treating hereditary diseases. Unfortunately, pathogenic variants are identified in 25%-60% of the patients. Current variant prioritization schemes rely on sequence features of variants, and are mostly oblivious to their tissue contexts. Here we report TRACEvar, a machine learning method for prioritizing pathogenic variants. The novelty of TRACEvar is in using 1,196 tissue-specific features, which provide tissue contexts for diseases, along with 117 sequence features of variants. TRACEvar was trained on ~68,000 mutations in 17 human tissues. TRACEvar random-forest models outperformed other variant prioritization tools both when assessed via cross-validation and when tested on data from 50 patients. TRACEvar also provided insight onto disease mechanism based on contributing features values. TRACEvar webserver is available at https://netbio.bgu.ac.il/TRACEvar/.

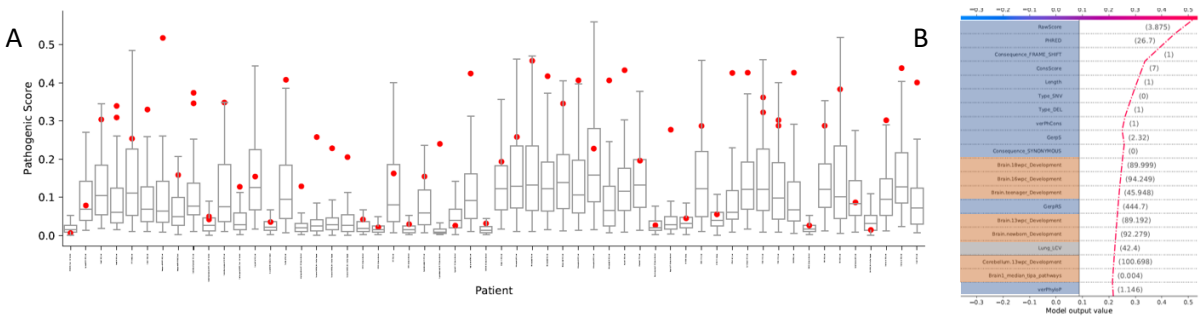

Fig.1. Prioritizing candidate variants in 50 patients.
A. The pathogenic scores of the variants per patient. Each box represents the pathogenicity score (Y axis) of the patient's candidate variants (X axis). A red dot marks the score of the true pathogenic variant in that patient. In 95% of the cases, the true pathogenic variant scored above the median score of that patient's variants, and in 56% of the cases it ranked among the top 10%.
B. Variant interpretation. SHAP decision plot for the true pathogenic variant of patient 28085 that suffers from mental retardation. The true pathogenic variant ranked 6 among the 240 candidate variants of that patient. The plot shows the top 20 most contributing features for the variant prediction in TRACEvar brain model. Notably, 7/20 were brain-specific features, 6 of which related to brain development. Sequence-based features were colored blue; features of brain tissue were colored orange; features of other tissue were colored gray.