

Using Machine Learning to Support Variant Interpretation in a Clinical Setting

Carmen Lai, B.S., Robert O’Connor, Ph.D., Scott Topper, Ph.D., FACMG, Jack Ji, Ph.D., FACMG, CGMB, Will Stedden, Ph.D., Julian Homburger, Ph.D., Jeroen Van den Akker, Ph.D., MB (ASCP)., Danny DeSloover, B.A.Sc., Alicia Zhou, Ph.D., Anjali Zimmer, Ph.D., Gilad Mishne, Ph.D.
Color Genomics, Burlingame, CA



Introduction

The American College of Medical Genetics and Genomics (ACMG) guidelines [1] provide an important framework for weighing and combining evidence in the clinical interpretation of variants. However, for some classes of variants—such as rare missenses, which are the majority of sequence variants—evidence can be limited or conflicting. For these variants, additional refinements to the guidelines have been proposed [2], and in some cases guidelines were found to be inconsistent with data [3].

Here, we present LEAP (Learning from Evidence to Assess Pathogenicity), a data-driven approach to combining evidence for variant interpretation using machine learning. LEAP produces transparent and evidence-based recommendations for the clinical interpretation of a variant, and is highly concordant with interpretation as performed by board-certified geneticists. We evaluate the predictive power of LEAP at different levels of evidence, and discuss its utility as an aid in the clinical interpretation process.

LEAP is novel in several ways. First, it is constructed and can be updated in a fully automated manner, and improves in accuracy as more data is available. Second, LEAP combines information commonly used in computational variant interpretation, such as functional prediction or conservation scores, with other evidence, such as population frequency, splicing impact, phenotypic information, and co-occurring variants. Lastly, in addition to high accuracy, the output is easy to understand as LEAP provides visualizations of individual evidence contribution to its recommendations.

Methods

Model: An L2-regularized logistic regression model was trained using Python’s “scikit-learn” library to output a predicted probability of pathogenicity for each variant (hereby referred to as predictions), which is used to derive an expected classification. Model weights and observed feature values were multiplied to determine the relative contribution of individual inputs (pathogenic vs. benign drivers) toward a recommendation for a given variant. Separately, a random forest classifier with 1,000 trees was trained and used to rank features in order of overall predictive importance.

Variants and Labels: A set of 2,563 rare (gnomAD MAF < 0.1%) missense variants classified as pathogenic (P, LP) or benign (B, LB) were used to train the model. Variants of uncertain significance (VUS) were excluded. Variants were detected by an NGS hereditary cancer panel in 24 loss-of-function cancer genes in which pathogenic variants have been associated with elevated risk for hereditary breast, ovarian, uterine/endometrial, colorectal, melanoma, pancreatic, prostate, and stomach cancer. These genes are *APC*, *ATM*, *BAP1*, *BARD1*, *BMPR1A*, *BRCA1*, *BRCA2*, *BRIPI*, *CDH1*, *CDKN2A*, *CHEK2*, *MLH1*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *PALB2*, *PMS2*, *PTEN*, *RAD51C*, *RAD51D*, *SMAD4*, *STK11*, and *TP53*. Variant classifications used as training labels were determined according to the ACMG 2015 guidelines for sequence variant interpretation [1], and approved by an American Board of Medical Genetics and Genomics certified medical geneticist.

Features: Quantitative and qualitative evidence considered during variant interpretation was used as model inputs. Evidence categories include functional predictions, evolutionary conservation scores, population and subpopulation variant frequencies, splicing impact, protein domain, pedigree phenotype, and co-variant information, as detailed in Table 1. Numeric features were standardized by centering at the median and scaling to the interquartile range. Categorical features were binarized, and pedigree and co-variant information were aggregated at a variant level. Missing values were filled using the most frequent value for numeric features, or filled with a “missing” label for categorical features.

Validation: Model performance as measured by area under the receiver operating characteristic curve (AUROC) was assessed using 10-fold cross-validated predictions. Performance of the model across different genes was also assessed using gene holdout cross-validated predictions, which were obtained for variants for each gene withheld from model training. Predictions from REVEL [4], a meta-predictor based on functional and conservation scores for rare missense pathogenicity prediction, were evaluated on the same validation set for AUROC comparison.

Conclusions

- LEAP is a variant interpretation tool that combines multiple categories of evidence for variant interpretation and significantly improves model performance (AUROC 97.9%) compared with algorithms using computational features only
- LEAP is extensible to multiple cancer loss-of-function genes, particularly those with high penetrance, demonstrating strong performance (AUROC 97.5%) even when a gene of interest is withheld from the training set
- Patient-level features like pedigree and co-variant information did not significantly improve overall AUC, but improved precision (positive predictive value) and recall (sensitivity)

Next Steps

Improvements: LEAP was trained using logistic regression, which provides visibility into individual evidence contribution to a given variant pathogenicity prediction, but represents contribution in a linear fashion. Alternatively, a non-linear (trees-based) approach could capture more nuanced patterns and provide more hierarchical rationale for a prediction, similar to that from a variant scientist’s thought process. Additionally, a more comprehensive feature set could be utilized (ClinVar consensus, literature content), and feature processing could be improved with more sophisticated missing value imputation (kNN).

Extensibility: Gene holdout cross-validation results show extensibility of the model to multiple loss-of-function genes for cancer. Extensibility to other health conditions, such as cardiovascular disease, has not been investigated. These conditions are less well understood than cancer, but as genetics starts to play a larger role, machine learning can serve as a tool to generate new criteria and increase efficiency in the variant interpretation process.

Results

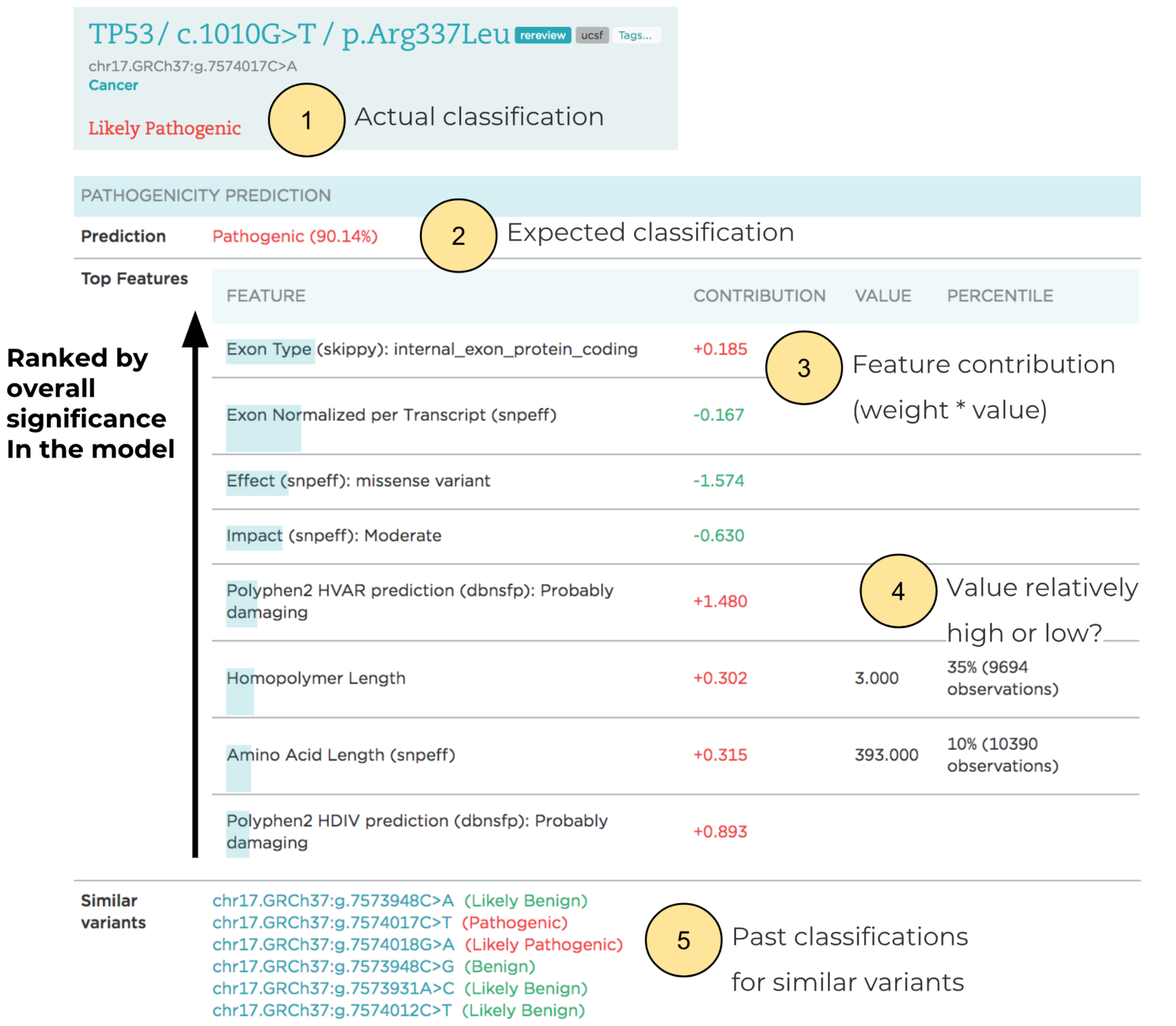
Table 1: Variant evidence inputs and significance

Feature inputs in order of overall significance. Order was determined by a random forest feature importance ranking which minimizes Gini impurity and optimizes for the highest quality (purest) decision tree splits.

Category	Source	Description
Functional predictor	Polyphen2-HVAR	Structural and functional impact prediction at amino acid level
Conservation	LRT	Amino acid constraint likelihood ratio test
Functional predictor	SIFT	Structural and functional impact prediction at amino acid level
Conservation	phastCons100way	Probability that nucleotide belongs to a conserved element
Conservation	GERP++	Rejected Substitution (RS) score compares observed substitutions across species with expected at random
Domain	Gene	Gene annotation
Population frequency	gnomAD	Summary data for African, Ashkenazi Jewish, East Asian, Finnish European, Latino, Non-Finnish European, and South Asian populations
Splicing impact	Skippy	Splicing impact prediction algorithm for exonic variants, enhancer and silencer elements
Domain	dbNSFP Interpro	Domain or conserved site of variant
Functional predictor	MutationTaster2	Structural and functional impact prediction at nucleotide level
Splicing impact	Alamut	4 RNA canonical sequences splicing impact predictions
Patient information	Color co-variant data	Variant co-occurrence with a known pathogenic variant
Patient information	Color health history data	Personal and family health history of various cancers

Figure 2: Model output visualization for use in a clinical setting

Visualization and explanation of model recommendations are shown as an aid for variant scientists during variant interpretation for clinical reporting. Contributing evidence features are ordered based on overall significance, and contribution magnitude and direction (pathogenic vs. benign driver) are displayed and color coded. Numeric feature percentiles with respect to the distribution observed in the training set are also shown for comparison. As additional context, classifications for similar variants (based on gene, chromosome, and exon) are listed.



References

1. Richards S, Aziz N, Bale S, Bick D, Das S, 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: 405–424.
2. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho Y-Y, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med. 2017;19: 1105–1117.
3. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018; doi:10.1038/gim.2017.210
4. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99: 877–885.

Figure 1: 10-fold cross-validation with varying levels of evidence

LEAP performance was assessed based on AUROC on 10-fold cross-validated predictions from models trained with different levels of evidence. Evidence categories include “computational” (functional prediction and conservation scores), “MAF” (population allele frequency), and “patient” (pedigree and co-variant data). LEAP trained with “computational only” evidence shows improved performance over REVEL (95.6% vs. 94.3%). The addition of population frequency and patient features further improves AUROC (97.9%).

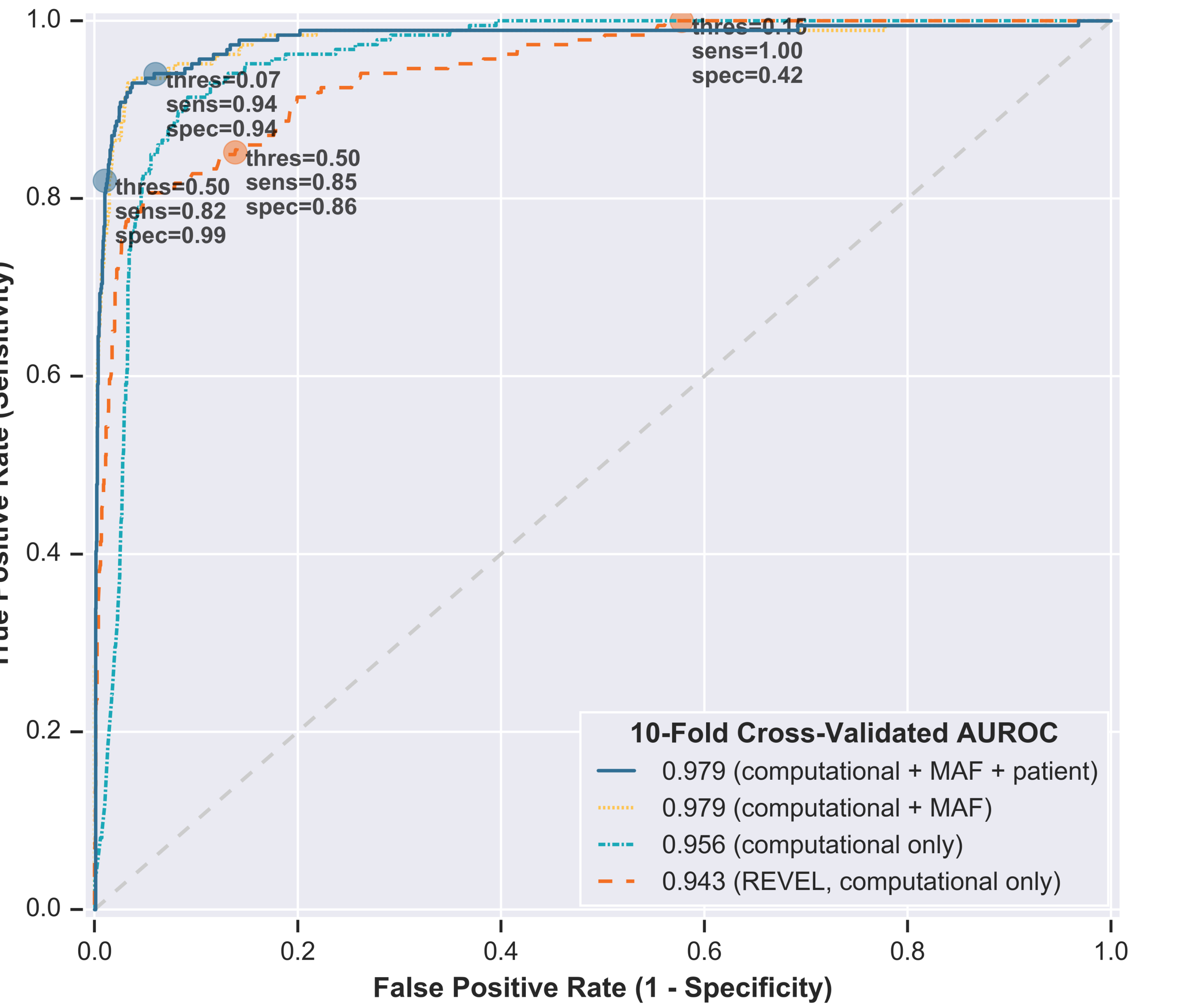


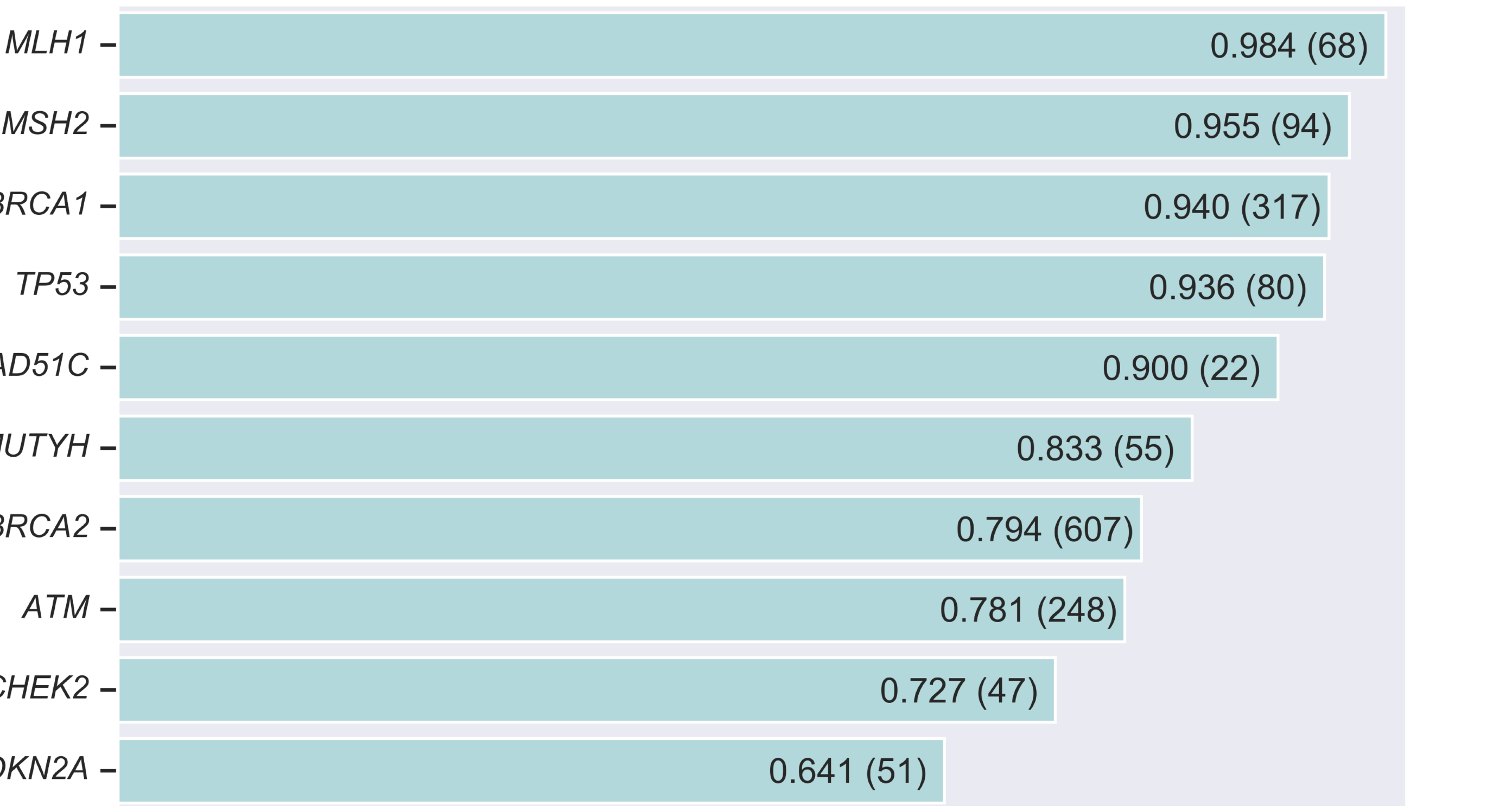
Table 2: Precision and recall for models with varying levels of evidence

True positive (TP), false positive (FP), false negative (FN), and true negative (TN) counts were determined using a 0.5 prediction cutoff for all model predictions. 10-fold cross-validated predictions were used to assess performance. LEAP at varying levels of evidence achieves higher precision compared with REVEL, although REVEL achieves somewhat higher recall. The addition of population frequency and patient information improves precision and recall for LEAP.

Validation Set	Precision TP/(TP+FP)	Recall TP/(TP+FN)	TP	FP	FN	TN
LEAP (computational + MAF + patient)	0.844	0.812	151	28	35	2349
LEAP (computational + MAF)	0.825	0.785	146	31	40	2346
LEAP (computational only)	0.627	0.704	131	78	55	2299
REVEL (computational only)	0.331	0.849	158	320	28	2057

Figure 3: Holdout cross-validation on loss-of-function cancer genes

AUROC breakdown by gene shows higher performance in genes with higher penetrance for cancer. Gene holdout cross-validated predictions achieved AUROC of 97.5% (compared with 97.9% with 10-fold cross-validated predictions). The weighted average across genes was 85.3%.



Gene Holdout AUROC (# Variants)
Genes with fewer than 5 pathogenic (P/LP) variants detected were excluded from this figure.