Machine learning models impact in reducing variants of uncertain significance in individuals undergoing genetic testing for gastrointestinal cancers

ID: P-095

Daniel E. Pineda-Alvarez, Brandie Heald, Sarah M. Nielsen, Edward D. Esplin, Britt A. Johnson, Laure Fresard, Yuya Kobayashi, Jason Reuter, Keith Nykamp, Swaroop Aradhya, Alexandre Colavin

Invitae Corporation, San Francisco, CA

Background & Aim

- The clinical classification of many missense variants is challenging due to limited available evidence. Consequently, they remain classified as variants of uncertain significance (VUS).
- VUS are at the core of healthcare disparities, as individuals from racial, ethnic and ancestral (REA) populations are underrepresented in large public genomic databases and medical literature and tend to receive more non-diagnostic results.
- We leveraged data from well represented populations to develop gene-specific machine learning (ML) models on the premise that disease variants will share similar molecular mechanisms.
- We evaluated the utility of these models in diagnostic panel testing for gastrointestinal and any type of cancer across REA groups.

Methods

- From 1/2022 to 5/2023, gene-specific ML algorithms were validated and integrated at Invitae® (Table 1).
- Evidence from the ML models were calibrated and incorporated into Sherloc, a semi-quantitative variant interpretation framework.[1]
- Evidence that met >80% NPV and PPV were incorporated during variant interpretation (Figure 1).
- VUS reduction rates were calculated, stratified by REA groups. Analyses of >20,000 patients were performed from random sampling of 20,000 patients with extrapolation. Measurement error was <2% variation by bootstrapping.

	Model	Type of evidence
External	Evolutionary model of variant effect (EVE)[2]	Conservation
	Multiplex assays of variant effects (MAVEs)[3–9]	Functional
	SpliceAI[10]	Splicing prediction
	Model	**Dataset leveraged**
Internal	Cellular evidence modelling[11]	Internally generated
	Gene specific engine[12]	Datasets from Refseq
	Molecular stability engine[12]	AlphaFold Protein Structures
	Multidimensional hotspots[13]	Known pathogenic variants from ClinVar
	Population frequency modeling[14]	gnomAD

Table 1. List of external and internally available models for use in Sherloc. Datasets from RefSeq include: amino acid sequence conservation, physicochemical conservation and homolog information

Results

- One or more applicable models were available for analyzing variants in 158 genes during the study period
- 187,767 individuals underwent diagnostic panel testing across our inherited metabolic disorders test menu including ~60,000 (32%) from an underrepresented REA group (Black, Hispanic, Asian)
- ~87,400 (47%) had ML evidence applied to ≥1 variant. Models contributed to the classification of at least 1 benign/likely benign (B/LB) variant in ~38,500 (21%) or pathogenic/likely pathogenic (P/LP) variant in ~1200 (0.6%) of individuals.

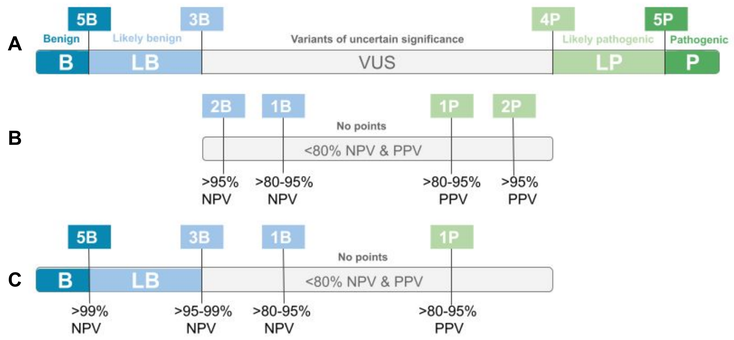

Figure 1. (A) Point thresholds for variant classification in Sherloc. Points awarded for non-population based ML model evidence **(B)** or population frequency modeling evidence **(C)** depending on negative predictive value (NPV) and positive predictive value (PPV) of the model. Population frequency modeling evidence has more benign weight, consistent with ACMG/AMP criteria[15]

- Across any, colorectal, gastric and pancreatic cancer history, a higher percentage of Black (23-30%), Asian (27-34%), and Hispanic (21-32%) individuals had a definitive classification that was dependent on ML evidence compared to White (14-23%, Table 2).

Cancer history	Clinician-reported REA group	Definitive classification n* (%)	B/LB n (%)	P/LP n (%)	*p*-value †
Any type	Any^	4187 (21)	4102 (21)	137 (1)	
	Asian	3069 (34)	3023 (34)	75 (0)	**9.80E-211**
	Black	3851 (29)	3800 (29)	89 (0)	**6.16E-134**
	Hispanic	5857 (32)	5742 (31)	227 (1)	**1.52E-225**
	White^	3569 (18)	3470 (17)	140 (0)	
Colorectal	Any^	2596 (16)	2519 (16)	87 (0)	
	Asian	261 (32)	258 (32)	3 (0)	**8.70E-48**
	Black	246 (25)	239 (24)	9 (1)	**1.23E-22**
	Hispanic	295 (21)	290 (21)	6 (0)	**1.62E-15**
	White	1321 (14)	1276 (13)	51 (1)	
Gastric	Any	817 (26)	799 (25)	38 (1)	
	Asian	55 (32)	55 (32)	1 (1)	**0.004**
	Black	62 (30)	61 (29)	1 (1)	**0.016**
	Hispanic	151 (30)	147 (29)	10 (2)	**0.001**
	White	365 (23)	356 (22)	14 (1)	
Pancreatic	Any	1300 (16)	1274 (15)	33 (0)	
	Asian	108 (27)	108 (27)	1 (0)	**3.49E-13**
	Black	136 (23)	135 (23)	2 (0)	**1.36E-10**
	Hispanic	125 (21)	124 (21)	2 (0)	**2.75E-07**
	White	665 (14)	645 (13)	22 (1)	

Table 2. ML model contribution to definitive variant classification stratified by cancer type and clinician reported race and ethnicity.

*some individuals might have >1 variant that was impacted by machine learning

† one-tailed, two-sample proportion test compared to individuals who were clinician-reported White

^analysis limited to sampled pool of 20,000 individuals

Conclusions

- By leveraging large publicly available data sets to create gene-specific ML algorithms for clinical assessment of variants, the resultant modeling evidence had an impact on a significant number of individuals undergoing gene panel testing for gastrointestinal cancers.
- Among individuals who had ≥1 variant with ML evidence applied towards its interpretation, over 20% resulted in definitive classifications.
- Validated use of gene-specific ML appears to provide more benefit during variant classification for individuals from traditionally underrepresented REA populations. ML models can assess biological and molecular factors in ways that are agnostic to population ancestry and, when systematically incorporated into clinical variant interpretation for inherited cancer syndromes, can narrow the gap in VUS rates among individuals from diverse population ancestries who undergo genetic testing.

References: **1**) Nykamp et al. Genet Med 19(10), 1105-1117 (2017) **2**) Frazer et al. 2021 Nature 599, 91–95, **3**) Findlay et al. Nature 562, 217–222 (2018) **4**) Richardson et al. Am J Hum Genet 108(3), 458-468 (2021) **5**) Jia et al. Am J Hum Genet 108(1), 163-175 (2021) **6**) Kato et al. PNAS 100(14), 8424-8429 (2003) **7**) Kotler et al. Mol Cell 71(1), 178-190 (2018) **8**) Giacomelli et al. Nat Genet 50(10), 1381-1387 (2018) **9**) Glazer et al. Am J Hum Genet 107(1), 111-123 (2020) **10**) Jaganathan et al., 2019, Cell 176, 535–548. **11**) Whitepaper, *in press* **12**) Manuscript, *in prep* **13**) Araya et al. Nat Genet. 48(2), 117-125 (2016) **14**) https://invit.ae/3k6Qm1G **15**) Richards et al. Genet Med. 17(5), 405-24 (2015) **Disclosures:** All authors are stockholders and employees of Invitae.