PREDICTING THE EFFECTS OF DNA VARIATIONS ON HUMAN HEALTH

Oct 28, 2020

Jay Park

WHAT IS DNA?

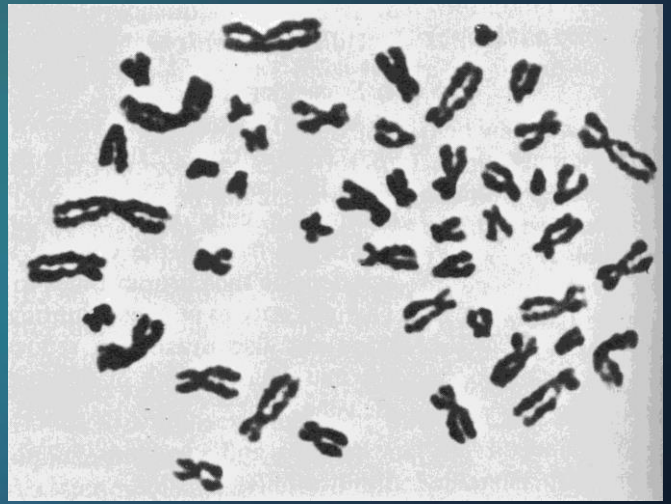
-Blueprints for life

-Humans have 46 pieces (usually in spaghetti form)

WHAT IS DNA?

-Blueprints for life

-Humans have 46 pieces (usually in spaghetti form)



CONDENSED FORM!

WHAT ARE GENETIC VARIATIONS? MUTATIONS? POLYMORPHISMS?

-Everyone’s DNA is very slightly different

Genetic Variation

(polymorphism / mutation)

WHAT ARE GENETIC VARIATIONS? MUTATIONS? POLYMORPHISMS?

-Everyone's DNA is very slightly different

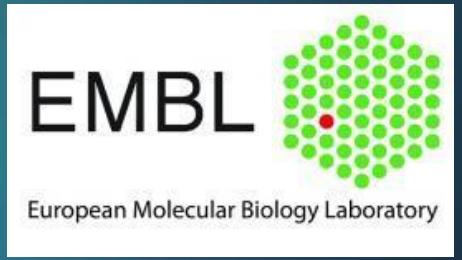
RESEARCH QUESTION

Can clinical significance of a variation be predicted purely through bioinformatic methods?

DATASET DESCRIPTION

ClinVar

ENSEMBL



354199 *Homo sapiens* genomic variations

IS THE VARIATION…

Benign / Likely benign

Uncertain

Pathogenic / Likely pathogenic

IS THE VARIATION…

RS 587784256
Gene: PAFAH1B1

Benign / Likely benign

Uncertain

Pathogenic / Likely pathogenic (1)

IS THE VARIATION…

RS 11203289
Gene: SDHB

Benign / Likely benign (7)

Uncertain (1)

Pathogenic / Likely pathogenic (1)

‘CLINICAL REVIEW STATUS’ TIERS

Reviewed by expert panel

Reviewed by multiple labs / No conflicts

Reviewed by multiple labs / Conflicting interpretations

Reviewed by single lab with justification

Only assertion provided (no justification)

RELIABILITY

‘CLINICAL REVIEW STATUS’ TIERS

57,663 rows

Reviewed by expert panel

Reviewed by multiple labs / No conflicts

Reviewed by multiple labs / Conflicting interpretations

Reviewed by single lab with justification

Only assertion provided (no justification)

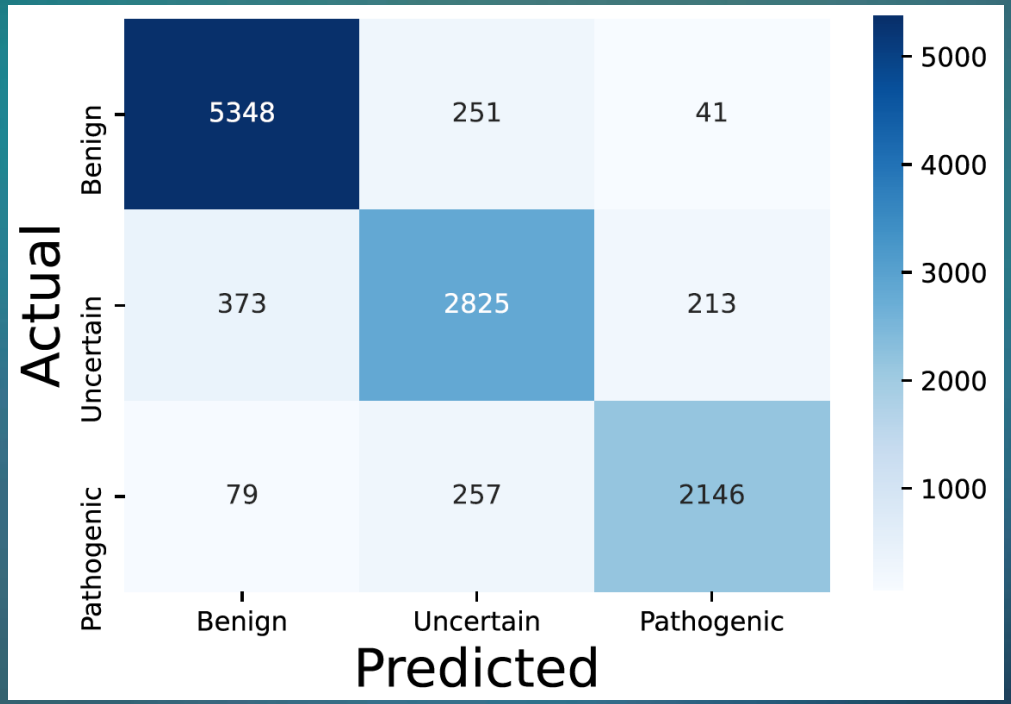
RELIABILITY

FEATURES FOR TRAINING

-Frequencies of variant in population (multiple studies)

-Various bioinformatic metrics estimating harmfulness

TEST RESULTS (XGBoost)



Accuracy: 0.89

	Precision	Recall
Benign	0.92	0.95
Uncertain	0.85	0.83
Pathogenic	0.89	0.86

APPLICATIONS

‘CLINICAL REVIEW STATUS’ TIERS

Reviewed by expert panel

Reviewed by multiple labs / No conflicts

Reviewed by multiple labs / Conflicting interpretations

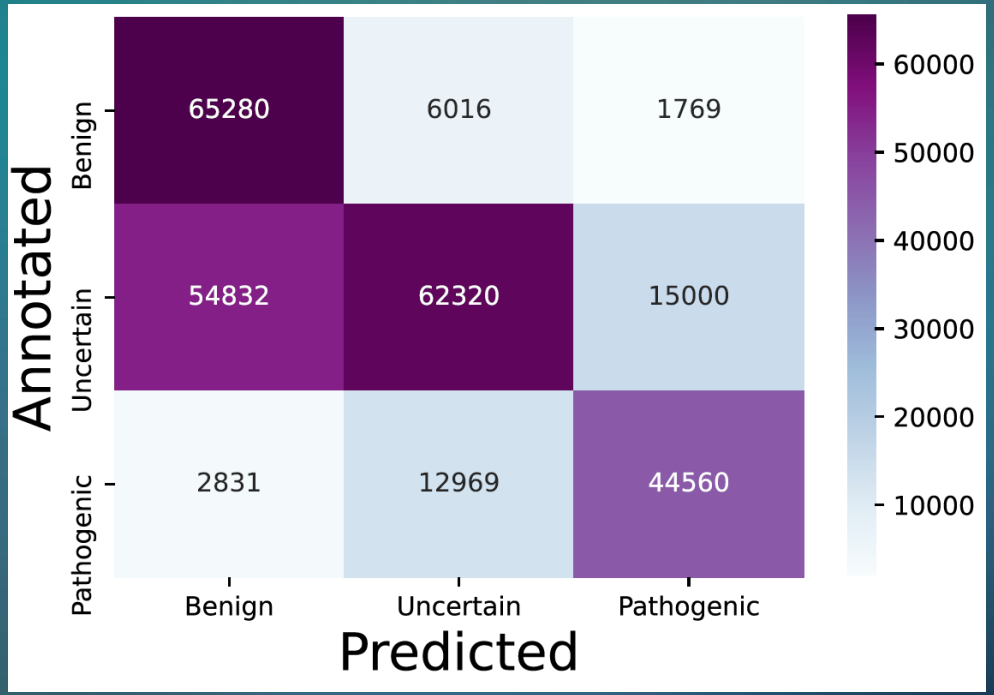
Reviewed by single lab with justification

Only assertion provided (no justification)

RELIABILITY

267,282 rows

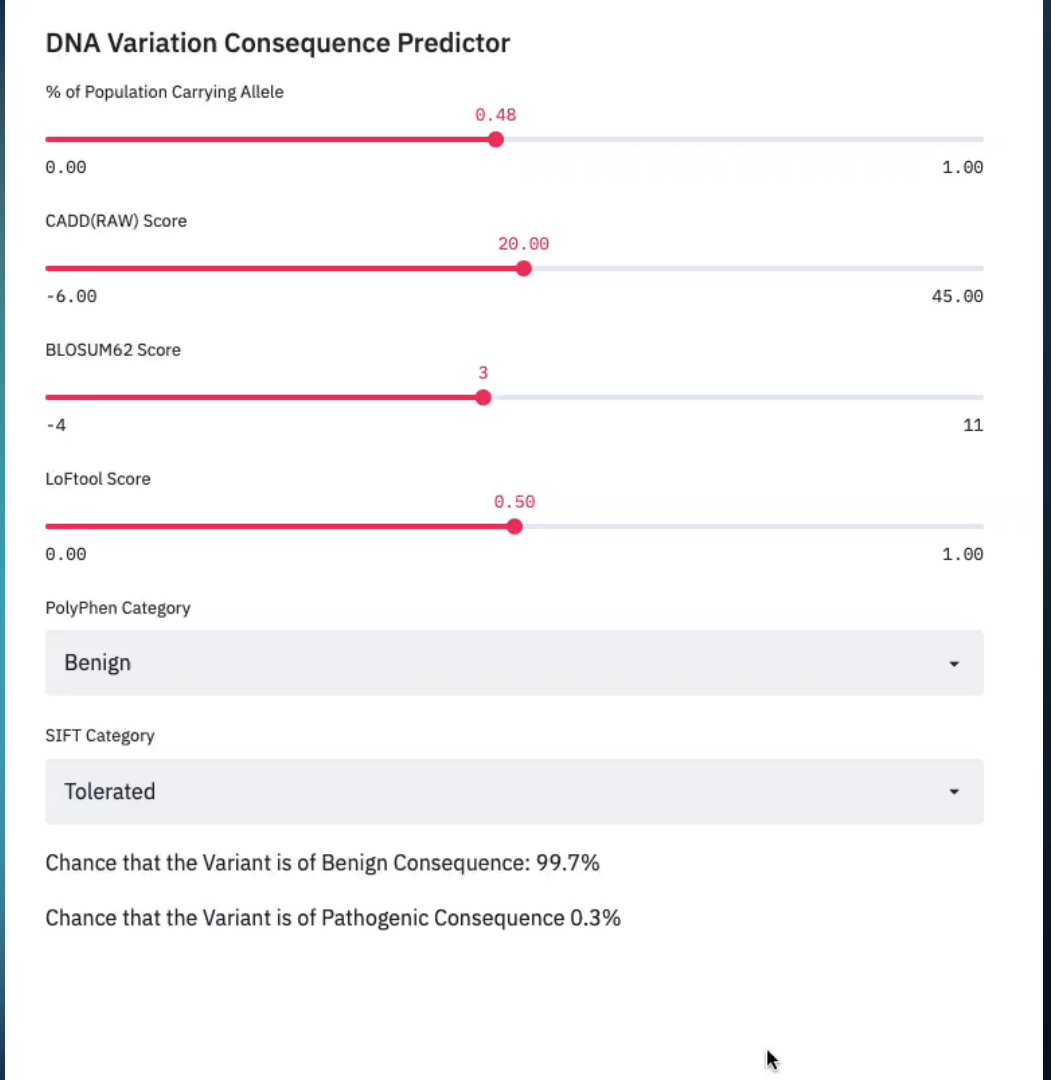
SINGLE-SUBMITTER ANNOTATIONS LARGELY AGREE WITH ALGORITHM



65% of annotations agree with model

Uncertain – Benign skew

TWEAKING MODEL PARAMETERS ON STREAMLIT



APPENDIX: FEATURES USED

-AF_ESP
-AF_EXAC
-AF_TGP
-CADD_RAW
-BLOSUM62
-LoFtool
-PolyPhen
-SIFT

APPENDIX: ACKNOWLEDGEMENTS

-European Bioinformatics Institute (@ EMBL)
-https://www.kaggle.com/kevinarvai/clinvar-conflicting
-ClinVar (@ NIH / NCBI)

-NHLBI GO Exome Sequencing Project
-Exome Aggregation Consortium
-1000 Genomes Project

-Rentzsch, Philipp et al. “CADD: predicting the deleteriousness of variants throughout the human genome.” Nucleic acids research vol. 47,D1 (2019): D886-D894. doi:10.1093/nar/gky1016
-Kircher, Martin et al. “A general framework for estimating the relative pathogenicity of human genetic variants.” Nature genetics vol. 46,3 (2014): 310-5. doi:10.1038/ng.2892
-Henikoff, S, and J G Henikoff. “Performance evaluation of amino acid substitution matrices.” Proteins vol. 17,1 (1993): 49-61. doi:10.1002/prot.340170108
-Fadista, João et al. “LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals.” Bioinformatics (Oxford, England) vol. 33,4 (2017): 471-474. doi:10.1093/bioinformatics/btv602
-Adzhubei, Ivan et al. “Predicting functional effect of human missense mutations using PolyPhen-2.” Current protocols in human genetics vol. Chapter 7 (2013): Unit7.20. doi:10.1002/0471142905.hg0720s76
-Ng, P C, and S Henikoff. “Predicting deleterious amino acid substitutions.” Genome research vol. 11,5 (2001): 863-74. doi:10.1101/gr.176601

THANK YOU!