APPLICATION NOTE

Machine-learning tools in TruSight™ Software Suite

Empower your variant analysis and interpretation in rare disease research with integrated machine-learning tools

illumina®

Introduction

Whole-genome sequencing (WGS) and whole-exome sequencing (WES) using next-generation sequencing (NGS) technologies provide a high-resolution view across the entire genome. Informatics solutions streamline analysis of the vast amounts of data produced by these methods, enabling discovery of variants associated with rare diseases. However, variant interpretation requires manual curation and scientific expertise, presenting a significant bottleneck when translating the raw sequencing data into meaningful, interpretable results efficiently.

In recent years, there has been increasing development and adoption of machine-learning techniques for NGS data analysis.[1] To this end, TruSight Software Suite incorporates several machine-learning and artificial intelligence (AI) tools to aid with variant interpretation and prioritization (Figure 1). This application note presents an overview of these advanced tools and how they can help push variant analysis into a new frontier.

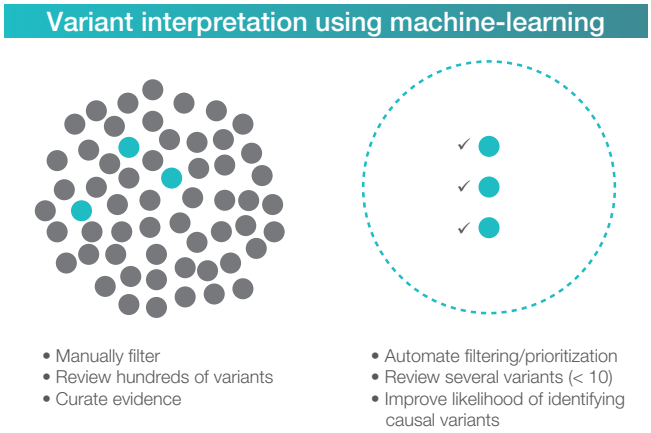

Figure 1: Variant interpretation using machine-learning—Machine-learning tools automate portions of variant analysis and interpretation, enabling users to focus on top candidates.

High-powered interpretation with machine learning

TruSight Software Suite includes "plug-and-play" machine-learning tools that can be run autonomously. By incorporating these automated prioritization tools, users can quickly filter out millions of variants to focus on candidate variants of interest for efficient and informed variant interpretation and curation.

SpliceAI

SpliceAI is a deep residual neural network that uses input genomic sequence to predict whether each position in a pre-messenger RNA (pre-mRNA) is a splice site (donor or acceptor). Splice donors and acceptors can be separated by large genomic distances.[2] In contrast to other tools that only consider short nucleotide sequences around exon–intron boundaries, SpliceAI evaluates 10K nucleotides of flanking sequence. This broad coverage enables accurate identification of noncoding mutations resulting from splicing errors that disrupt normal transcription and translation (Figure 2). For variants annotated with a SpliceAI score, SpliceAI predicts the likelihood of gain or loss of a splice donor or acceptor. This score ranges from 0.0 to 1.0, with values closer to 1.0 indicating higher confidence of an event.

PrimateAI

PrimateAI is a deep learning network developed for highly accurate classification of mRNA missense variants. Training the network on existing databases of human variants with limited coverage of the exome could lead to interpretation biases. To avoid bias, PrimateAI has been trained on a data set of human variants and > 300K unique missense variants collected from six nonhuman primate species, increasing overall performance in classifying pathogenic alleles (Figure 3).[3] For variants annotated with a PrimateAI score, PrimateAI predicts the likelihood of the variant of having a pathogenic effect. This score ranges from 0.0 to 1.0, with values closer to 1.0 indicating higher confidence of pathogenicity.

Emedgene

Emedgene, an Illumina partner, has developed a genomics AI engine that automates portions of the variant interpretation process to streamline data analysis. During variant filtering in TruSight Software Suite, users can filter small variants, including single nucleotide variants (SNVs) and insertions/deletions (indels) with the Emedgene tool for quick prioritization. Emedgene automatically filters and

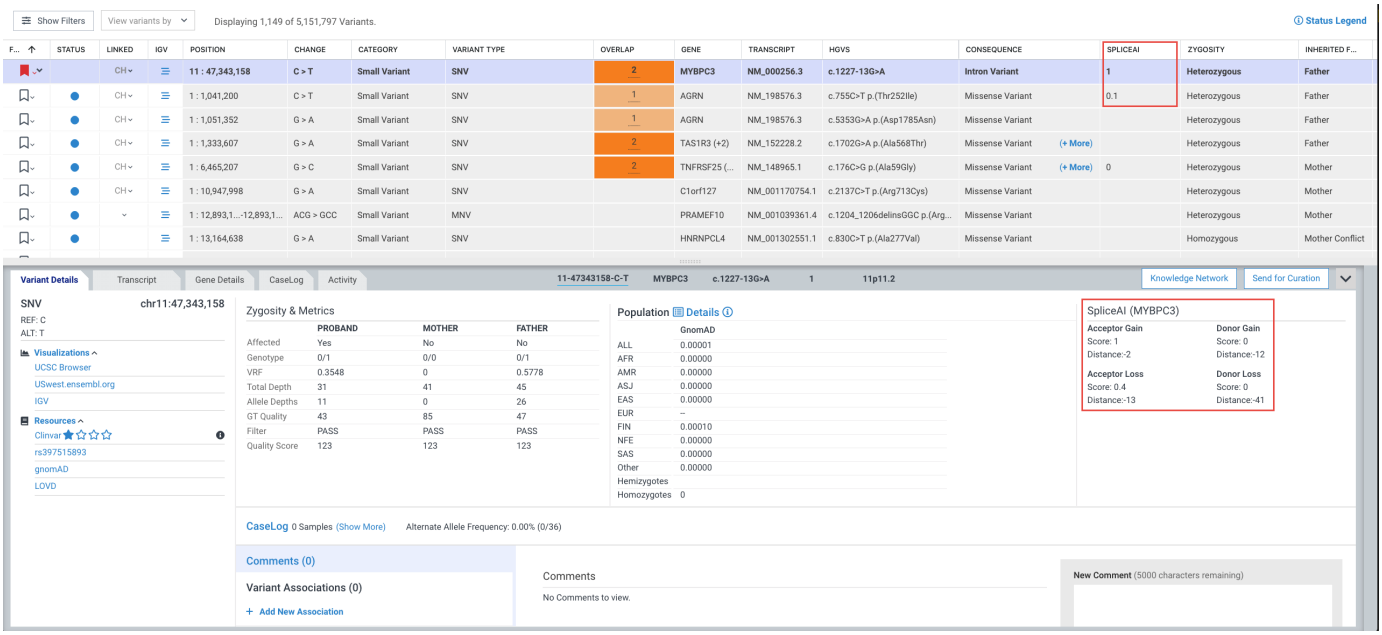

Figure 2: Evaluate flanking sequence with SpliceAI—TruSight Software Suite includes SpliceAI for evaluating 10,000 nucleotides of flanking sequence to identify noncoding mutations resulting from splicing errors. SpliceAI scores (displayed in the variant grid and Variant Details tab) predict the likelihood of a splicing event, with values closer to 1.0 indicating higher confidence of prediction.

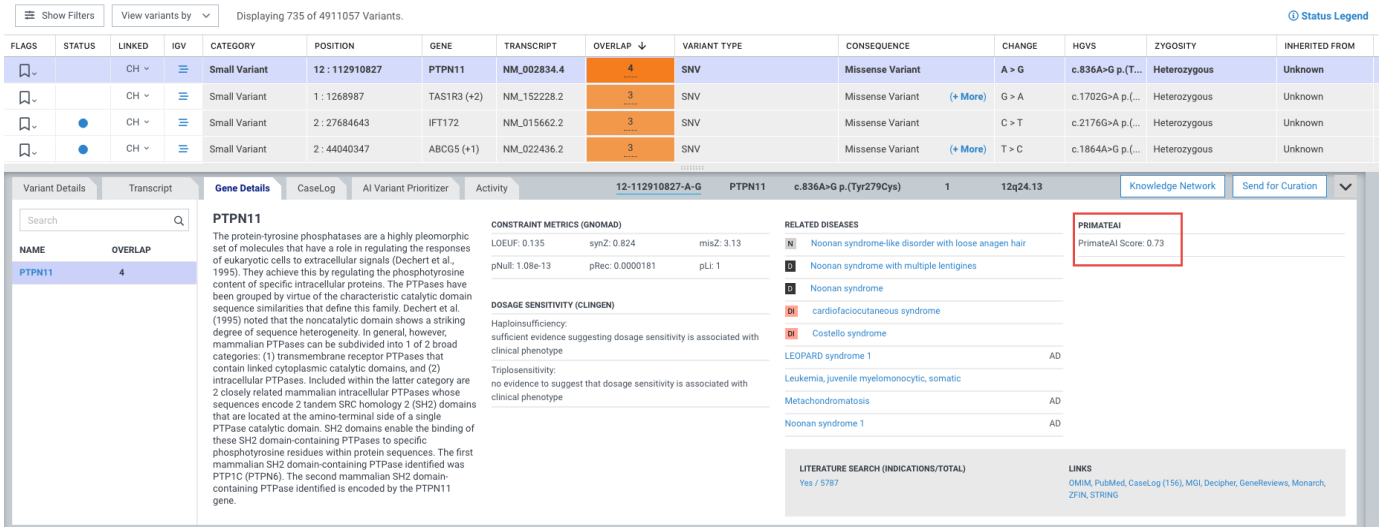

Figure 3: Classify missense variants with PrimateAI—TruSight Software Suite includes PrimateAI for highly accurate classification of mRNA missense variants, based on a training data set of human variants and > 300K unique missense variants collected from six nonhuman primate species. The Primate AI score for a variant (displayed in the Gene Details tab) predicts the likelihood of a variant being pathogenic, with values closer to 1.0 indicating higher confidence of prediction.

ranks variants according to a proprietary algorithm, presenting the top variants, identified as either "candidate" or "most likely," for review and evaluation (Figure 4).

Emedgene applies natural language processing (NLP) to various data sources to generate a Knowledge Graph for each ranked variant. Users can explore the Knowledge Graph to understand how and why a variant is prioritized by reviewing the variant details, the respective gene in which it occurs, and associations with "clinical indications," including known diseases and both confirmed and unconfirmed phenotypes (Figure 5). Users can follow links to supporting evidence from the scientific literature and online databases for more information.

Example use case with Emedgene

In 2018, a laboratory analyst in the Illumina Clinical Services Laboratory received a request for WGS variant analysis for two unaffected parents and a male proband with a phenotype that included neurodevelopmental delay. Initially, manual variant filtering, prioritization, and interpretation resulted in a report with no variants identified as having associations with known diseases.

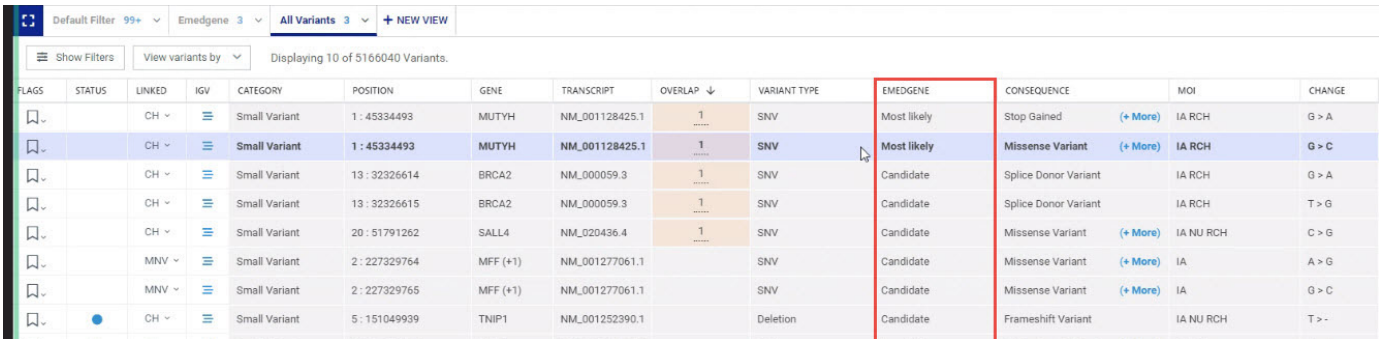

Figure 4: Filter and prioritize small variants with Emedgene—Applying the Emedgene filter set runs an automated variant filtering and ranking algorithm that returns the top variants ranked as either "candidate" or "most likely," based on the likelihood that they will solve a case.

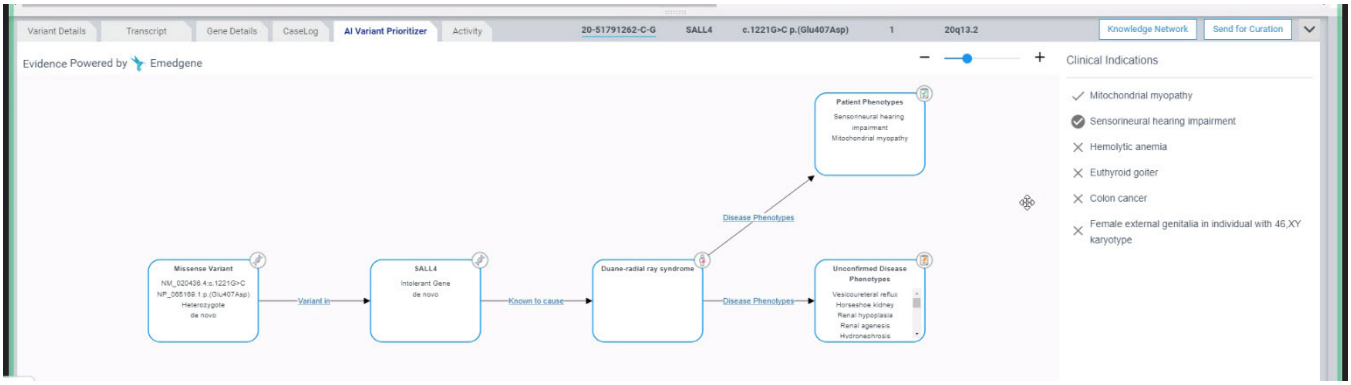

Figure 5: Knowledge Graph—The Emedgene Knowledge Graph displays information at a glance that helps users understand how and why a variant is prioritized, the variant details, the respective gene in which it occurs, and associations with "clinical indications," including known diseases and both confirmed and unconfirmed phenotypes. Knowledge Graph includes links to external online databases and the scientific literature with supporting information for the indicated associations.

After reanalysis with the Emedgene tool a year later, a variant in the thousand and one (TAO) amino acid kinase 1 (*TAOK1*) gene was prioritized. Emedgene automatically applied NLP to online databases and found a research article published in the *American Journal of Human Genetics* describing *de novo* variants in *TAOK1* associated with neurodevelopmental disorders.[4] Review and curation of the called variant and corresponding research article resulted in the analyst issuing an amended report (Table 1). Importantly, this result highlights the potential of machine-learning tools to enable routine, automated reanalysis of unsolved cases, which is not feasible by manual methods.

Table 1: Proband variant interpretation

Variant of interest	Interpretation
TAOK1 Small Variant (c.557C>T; p.Pro186Leu) Heterozygous	The c. 557C>T; p.Pro186Leu missense variant is classified as likely pathogenic

Summary

TruSight Software Suite incorporates "plug-and-play" machine-learning tools that can be run autonomously, without the need for bioinformatics or programming expertise. With these tools, users can quickly filter out millions of variants to focus on the top, candidate variants of interest for efficient and informed variant interpretation and curation.

Learn more

TruSight Software Suite, illumina.com/trusight-software-suite

References

1. Schmidt B, Hildebrandt A. Deep learning in next-generation sequencing. *Drug Discov Today.* 2020;S1359-6446(20)30415-3.
2. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell.* 2019;176(3):535–548.
3. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet.* 2018;50(8):1161–1170.
4. Dulovic-Mahlow M, Trinh J, Kandaswamy KK, et al. De novo variants in TAOK1 cause neurodevelopmental disorders. *Am J Hum Genet.* 2019;105(1):213–220.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-00008 v1.0..