Machinelearning tools in TruSight[™] Software Suite

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

illumına®

For Research Use Only. Not for use in diagnostic procedures.

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.

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.

SpliceAl

SpliceAl 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 exonintron boundaries, SpliceAl evaluates 10K nucleotides of flanking sequence. This broad coverage enables accurate identification of noncoding mutations



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

resulting from splicing errors that disrupt normal transcription and translation (Figure 2). For variants annotated with a SpliceAl score, SpliceAl 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.

PrimateAl

PrimateAl 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, PrimateAl 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 PrimateAl score, PrimateAl 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 Al 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

^	STATUS	LINKED	IGV	POSITION	0	HANGE	CATEGORY	VARIANT TY	PE	OVERLAP	GENE	TRANSCRIPT	HGVS	CONSEQUENCE		SPLICEAI	ZYGOSITY	INHERITED
~		CH~	=	11 : 47,343,1	158 C	> T	Small Variant	SNV		2	MYBPC3	NM_000256.3	c.1227-13G>A	Intron Variant		1	Heterozygous	Father
٦.		сн∽	-	1:1,041,200	c	> T	Small Variant	SNV		1	AGRN	NM_198576.3	c.755C>T p.(Thr252Ile)	Missense Variant		0.1	Heterozygous	Father
۹. ۲		CHV	=	1:1,051,352		> A	Small Variant	SNV		1	AGRN	NM_198576.3	c.5353G>A p.(Asp1785Asn)	Missense Variant			Heterozygous	Father
4~	•																	
٦-	•	CH~	=	1:1,333,607	G	> A	Small Variant	SNV			TAS1R3 (+2)	NM_152228.2	c.1702G>A p.(Ala568Thr)	Missense Variant	(+ More)		Heterozygous	Father
٦-	•	CH ~	=	1 : 6,465,207	G	> C	Small Variant	SNV		2	TNFRSF25 (NM_148965.1	c.176C>G p.(Ala59Gly)	Missense Variant	(+ More)	0	Heterozygous	Mother
٦.	•	CH ~	=	1:10,947,99	8 G	> A <	Small Variant	SNV			C1orf127	NM_001170754.1	c.2137C>T p.(Arg713Cys)	Missense Variant			Heterozygous	Mother
٦.	•	v	=	1:12,893,1	12,893,1 A	CG > GCC	Small Variant	MNV			PRAMEF10	NM_001039361.4	c.1204_1206delinsGGC p.(Arg	Missense Variant			Heterozygous	Mother
J.	•		=	1:13,164,63	8 G	> A	Small Variant	SNV			HNRNPCL4	NM_001302551.1	c.830C>T p.(Ala277Val)	Missense Variant			Homozygous	Mother Co
_																		
EFF: C 社:: Turner: CUSS Browser USS Browser IGV Resources へ Climat 会会会。 Climat 会会会。 Climat 会会会。 Climate 会会会。 Climate 会合会。 Climate 会合合合。 Climate 合合合合合。 Climate 合合合合合合。 Climate 合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合合				Affected Genotype VRF Total Depth Allele Depths GT Quality Filter	PROBAN Yes 0/1 0.3548 31 11 43 PASS	4D M4 No 0/4 41 0 85 PA		FATHER No 0/1 0.5778 45 26 47 PASS	ALL AFR AMR ASJ EAS EUR FIN	GnomAD 0.00000 0.00000 0.00000 0.00000 0.00010				Si Di Ad	cceptor Gain core: 1 stance:-2 cceptor Loss core: 0.4 stance:-13	Donor Gain Score: 0 Distance:-12 Donor Loss Score: 0 Distance:-41		
			Quality Score 123 123				123	123 NPE 0.0000 SAS 0.0000 Other 0.0000 Hemizygotes Homogygotes 0										
					CaseLog 0 Comments		ow More) Alterna	te Allele Frequer	icy: 0.00% (0/36)									
		Variant Associations (0)					Comments						New Comment (5000 characters remaining)					

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.

LAGS	STATUS	LINKED	IGV	CATEGORY	POSITION	GENE	TRANSCRIPT	OVERLAP 🗸	VARIANT TYPE		CONSEQUENCE		CHANGE	HGVS	ZYGOSITY	INHERITED F
۵.		CH ~	≡	Small Variant	12 : 112910827	PTPN11	NM_002834.4		SNV		Missense Variant		A > G	c.836A>G p.(T.	Heterozygous	Unknown
۵.		CH ~	=	Small Variant	1 : 1268987	TAS1R3 (+2) NM_152228.2	3	SNV		Missense Variant	(+ More)	G > A	c.1702G>A p.(.	Heterozygous	Unknown
۵.	•	CH ~	∍	Small Variant	2:27684643	IFT172	NM_015662.2	3	SNV		Missense Variant		C > T	c.2176G>A p.(.	Heterozygous	Unknown
۵.	•	CH ~	=	Small Variant	2:44040347	ABCG5 (+1)	NM_022436.2	3	SNV		Missense Variant	(+ More)	T > C	c.1864A>G p.(.	Heterozygous	Unknown
Variant	t Details	Transcri	ot	Gene Details	CaseLog Al Varian	Prioritizer A	activity	12-1129108	27-A-G PTPN11	c.836A>G	p.(Tyr279Cys)	1	12q24.13	к	nowledge Network	Send for Curation
Search Q PTPN11				CONSTRAINT METRIC	S (GNOMAD)		RELATED DISE	ASES			PRIMATEAI					
AME	ME OVERLAP The protein-tyrosine phosphatases are a highly pleomo set of molecules that have a role in regulating the respo		the responses	LOEUF: 0.135 synZ: 0.824 misZ: 3			N Noonar	n syndrome-like disorde	r with loose and	agen hair	PrimateAl Score: 0.73					
			1995). They achieve this by regulating the phosphotyrosine			pNull: 1.08e-13 pRec: 0.0000181 pLi: 1			D Noonar	n syndrome with multipl	e lentigines					
				content of specific intracellular proteins. The PTPases have been grouped by virtue of the characteristic catalytic domain						D Noonar	n syndrome					
					es that define this family. he noncatalytic domain s		DOSAGE SENSITIVITY (CLINGEN) Haploinsufficiency:			Di cardiofaciocutaneous syndrome						
				degree of sequence heterogeneity. In general, however, mammalian PTPases can be subdivided into 1 of 2 broad			sufficient evidence suggesting dosage sensitivity is associated with			DI Costello syndrome						
			categories: (1) trans	smembrane receptor PTF	ases that	clinical phenotype		LEOPARD syndrome 1 A Leukemia, juvenile myelomonocytic, somatic								
				intracellular PTPase	plasmic catalytic domain es. Included within the lat	ter category are	Triplosensitivity: no evidence to sugg					est that dosage sensitivity				
				sequences encode	ammalian intracellular P1 2 tandem SRC homology	2 (SH2) domains	clinical phenotype			Metachondro	matosis		AD			
				PTPase catalytic do	the amino-terminal side o omain. SH2 domains ena	le the binding of				Noonan syndr	rome 1		AD			
thee SH2 domain-containing PTPases phosphoryosine residues within protei- mammalian SH2 domain-containing PT PTPTIC (PTPNS). The second mammali containing PTPase identified is encode gene.				sidues within protein seq omain-containing PTPase ne second mammalian SF	uences. The first identified was 12 domain-				LITERATUR Yes / 5787	RE SEARCH (INDICATION	S/TOTAL)		LINKS OMIM, PubMed, Ca ZFIN, STRING	aseLog (156), MGI, Decipi	her, GeneReviews, Monarch,	

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.

-	au 10200 - 2	1.87 3	SK8 12	100	0. 0.000000000000000000000000000000000										
=	Show Filters	View var	iants by	 Displaying 	10 of 5166040 Variants.										
LAGS	STATUS	LINKED	IGV	CATEGORY	POSITION	GENE	TRANSCRIPT	OVERLAP 🕹	VARIANT TYPE	EMEDGE	NE	CONSEQUENCE		MOI	CHANGE
۵.		CH ~	=	Small Variant	1:45334493	MUTYH	NM_001128425.1	1	SNV	Most like	ely	Stop Gained	(+ More)	IA RCH	G > A
□.		CH ~	=	Small Variant	1 : 45334493	MUTYH	NM_001128425.1	1	SNV	Most lik	ely	Missense Variant	(+ More)	IA RCH	G > C
۵.		CH ~	=	Small Variant	13:32326614	BRCA2	NM_000059.3	.1	SNV	Candida	ite	Splice Donor Variant		IA RCH	G > A
$\square_{}$		CH ~	=	Small Variant	13:32326615	BRCA2	NM_000059.3	_1	SNV	Candida	ite	Splice Donor Variant		IA RCH	T > G
۵.		CH ~	=	Small Variant	20:51791262	SALL4	NM_020436.4	1	SNV	Candida	ite	Missense Variant	(+ More)	IA NU RCH	C > G
		MNV ~	=	Small Variant	2:227329764	MFF (+1)	NM_001277061.1		SNV	Candida	ite	Missense Variant	(+ More)	IA	A > G
۵.		MNV ~	-	Small Variant	2:227329765	MFF (+1)	NM_001277061.1		SNV	Candida	ite	Missense Variant	(+ More)	IA	G > C
Π.	•	CH ~	=	Small Variant	5:151049939	TNIP1	NM_001252390.1		Deletion	Candida	ite	Frameshift Variant		IA NU RCH	T>-

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
<i>TAOK1</i> Small Variant	The c. 557C>T; p.Pro186Leu
(c.557C>T; p.Pro186Leu)	missense variant is classified
Heterozygous	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.

illumina®

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

© 2021 Illumina, Inc. All rights reserved. All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html. M-GL-00008 v1.0..

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.
- 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.
- 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.