

Data-Driven Diagnosis of Rare Genetic Disorders

- Toward Medical Device Development

Sungwon Jung, Ph.D

Department of Genome Medicine and Science, Gachon University College of Medicine

Gachon Institute of Genome Medicine and Science, Gachon University Gill Medical Center

Dec 19, 2023 @DTMBIO2023

What are Rare Genetic Disorders?

Definition

- Diseases caused by genetic abnormalities among diseases in which prevalence < 20,000 or diagnosis is difficult (in Korea)
 - Varies by countries
(US: < 200,000, Japan: < 50,000, Europe: Ratio < 1/2,000, etc.)

Characteristics

- **Genetic**
 - Hereditary: 80%
 - Pediatric: > 50%
- **High mortality**
 - 35% of deaths within the first year of life
 - 30% die within five years of life
- **Diversity: More than 7,000 diseases reported**
 - Diversity in phenotypes and genotypes



That's more than
2.8 million Canadians.



...that's 3.5
million
people in the
UK alone.



Affecting 5 ~ 10% of population

Difficulty in diagnosis



- Due to rarity, diversity, genetic natures
- Visits 7 MDs on average until diagnosis
- 5 to 7 years on average until correct diagnosis

Value of diagnosis

- Gives answer to patients with the cause of disease
- Potential chances for treatment
- Genetic counseling to patients and parents
- Leads to new drug R&D
 - Personalized anti-sense oligonucleotide (ASO), etc.
 - Life-long treatment

Diagnosis of Rare Genetic Disorders



선천성 근무력증 11살 보경이가 테어나 처음 걷게 된 과정

2003년 희귀 유전성 질환인 선천성 근무력증으로 출생

두 살 때 근육 조직 검사 결과 근무력증 진단

지난해 열 살 때까지 줄곧 눕거나 누군가 앉혀주던 앉아 생활해옴

서울대병원, 미국 워싱턴대 의대 유전체 분석 연구소에 검사 의뢰



근육질환 관련 DOK7 유전자 변이 발견

올해 초부터 신경과 근육 연결 작용 활성 신경 물질(아세틸콜린) 강화제 두어

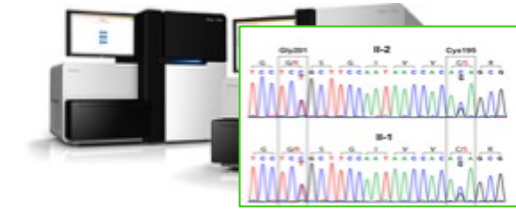
6월부터 조금씩 일어서기 시작

10월 현재 화장실 걸어서 가고 계단 오르기 시작

자료: 서울대병원 손이신경과



Clinical observation of patient's abnormal phenotypes



Identifying patient's genetic variations



Listing candidate ★ diseases



Previously reported > 7,000 disorders
Phenotypes
Causal genes



Gene X

Prioritizing potentially ★ causative pathogenic genetic variants

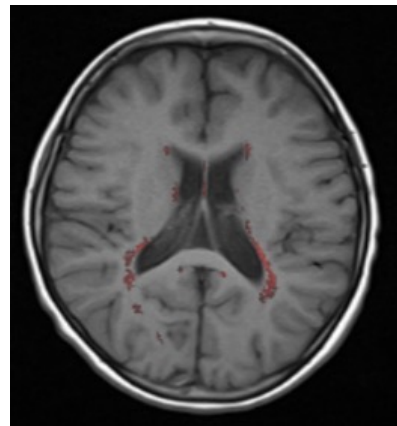
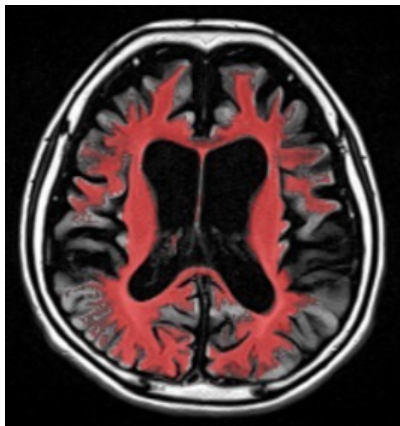
Final diagnosis of patient

Successful diagnosis is significantly dependent on clinical and technical abilities.

Difficulties in Diagnosis of Rare Genetic Disorders

- ◆ Many different diseases with not-quite-clear phenotype differences
 - Around 7,000 known rare genetic disorders
 - A physician cannot be familiar with all of them.

- ◆ Phenotypic heterogeneity can happen in single disease.

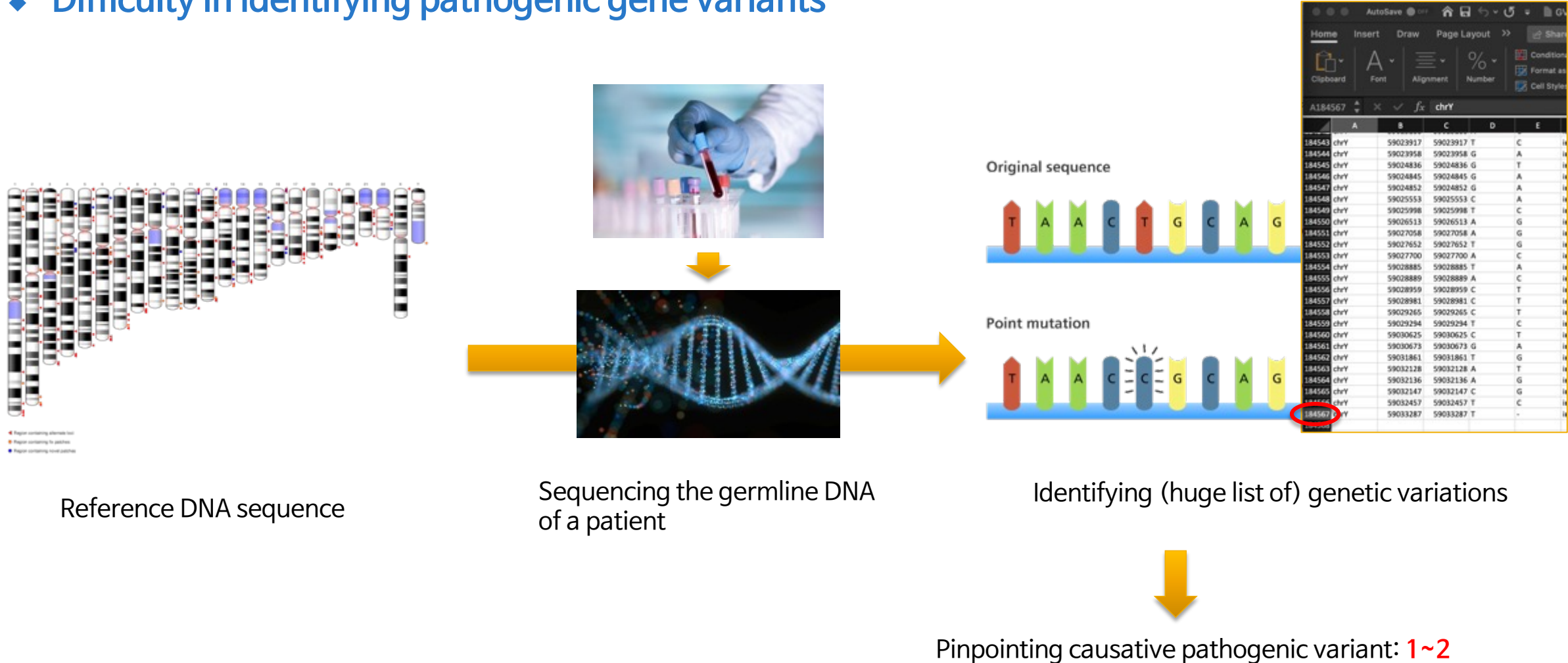


Ceroid lipofuscinosis, neuronal, 6 (CLN6)

Same disease, very different phenotypes

Difficulties in Diagnosis of Rare Genetic Disorders

◆ Difficulty in identifying pathogenic gene variants



Original sequence



Point mutation



	A	B	C	D	E
184543	chrY	59023917	59023917	T	C
184544	chrY	59023958	59023958	G	A
184545	chrY	59024836	59024836	G	T
184546	chrY	59024845	59024845	G	A
184547	chrY	59024852	59024852	G	A
184548	chrY	59025553	59025553	C	A
184549	chrY	59025998	59025998	T	C
184550	chrY	59026513	59026513	A	G
184551	chrY	59027058	59027058	A	G
184552	chrY	59027652	59027652	T	G
184553	chrY	59027700	59027700	A	C
184554	chrY	59028885	59028885	T	A
184555	chrY	59028889	59028889	A	C
184556	chrY	59028959	59028959	C	T
184557	chrY	59028981	59028981	C	T
184558	chrY	59029265	59029265	C	T
184559	chrY	59029294	59029294	T	C
184560	chrY	59030625	59030625	C	T
184561	chrY	59030673	59030673	G	A
184562	chrY	59031861	59031861	T	G
184563	chrY	59032128	59032128	A	T
184564	chrY	59032136	59032136	A	G
184565	chrY	59032147	59032147	C	G
184566	chrY	59032457	59032457	T	C
184567	chrY	59033287	59033287	T	-

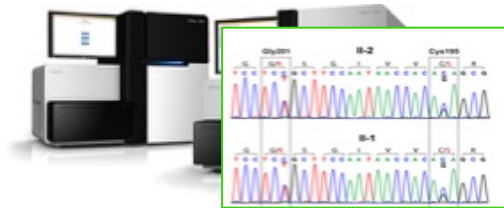
Typical Diagnosis Steps of Rare Genetic Disorders

Step 1: Clinical diagnosis

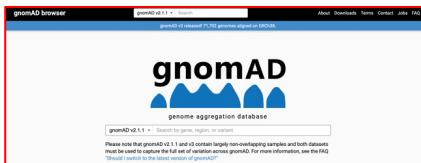


- Identifying abnormal phenotypes
- (Listing candidate diseases)

Step 2: Identifying rare genetic variants



Identifying genetic variations



Filtering genetic variants with low VAF

Step 3: Prioritizing pathogenic genetic variants

Possible large variation in diagnosis

	Benign		Pathogenic		
	Strong	Supporting	Supporting	Moderate	Strong
Population data	MAF is too high for disorder (RA > 1%) OR observed in controls inconsistent with disease penetrance BS2		Absent in population databases P02		Prevalence in affecteds historically increased over controls P04
Computational and predictive data	Multiple lines of computational evidence suggest no impact on gene/ligand product BP4 Missense in gene amino acid only noncoding lesion ClinVar BP1 Silent variant with non-predicted splice impact BP7 In-frame indels in repeat/short tandem repeats BP3	Multiple lines of computational evidence suggest a deleterious effect on the gene/ligand product P03	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen P05 Frameshift length changing variant P04	Same amino acid change as an established pathogenic variant P01	Predicted null variant in a gene where LOF is a known mechanism of disease P05
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with location of benign missense variants and path. missense common P02	Missional null spot in well-established functional domain without benign variation P01	Well-established functional studies show a deleterious effect P03
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in highly affected family members P01	Increased segregation data	
De novo data			De novo (without parental & maternal confirmed) P02	De novo (paternity and maternity confirmed) P02	
Allelic data	Observed in homo with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in homo with pathogenic variant P04		
Other databases	Reputable source archival shared data - benign BP5		Reputable source - pathogenic P05		
Other data	Found in case with an alternate cause		Patient's phenotype or PTA highly specific for gene P04		

	A	D	E	F	G	H	AJ	AK	AL	AM	AN	AO
1 Gene.refGene	chr	Start	End	Ref	Alt		SIFT_score	SIFT_pred	Polyphen2_H	Polyphen2_H	Polyphen2_H	Polyphen2_H
2 GJB4	chr1	35226964	35226964	G	A		0.06	T	0.993	D	0.706	P
3 FPGT-TNNI3	chr1	74716436	74716436	C	G		0.01	D	1	D	0.996	D
4 PPOX	chr1	161138854	161138854	C	G		0.04	D	0.997	D	0.944	D
5 FCGR2A	chr1	161483723	161483723	G	A		-999	.	-999	.	-999	.

- Prioritization based on diagnosis guidelines (e.g., ACMG)
 - Specific implementation of each guideline step is mandatory
 - Computational prediction of pathogenicity is not good enough

Step 4: Final diagnosis

OMIM® Online Mendelian Inheritance in Man
An Online Catalog of Human Genetic Disease
Updated October 30, 2019

256000
LEIGH SYNDROME; LS **질현형**

Alternative titles/synonyms
NECROTIZING ENCEPHALOPATHY, INFANTILE SUBACUTE, OF LEIGH; SNE

Other entities represented in this entry:
LEIGH SYNDROME DUE TO MITOCHONDRIAL COMPLEX I DEFICIENCY, INCLUDED
LEIGH SYNDROME DUE TO MITOCHONDRIAL COMPLEX II DEFICIENCY, INCLUDED
LEIGH SYNDROME DUE TO MITOCHONDRIAL COMPLEX III DEFICIENCY, INCLUDED
LEIGH SYNDROME DUE TO MITOCHONDRIAL COMPLEX IV DEFICIENCY, INCLUDED
LEIGH SYNDROME DUE TO MITOCHONDRIAL COMPLEX V DEFICIENCY, INCLUDED

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus number
2q35	Leigh syndrome	256000	Mt, AR	3	BCSL1 603647
5p15.33	Leigh syndrome	256000	Mt, AR	3	SDHA 603857
9q34.2	Leigh syndrome, due to COX IV deficiency	256000	Mt, AR	3	SLRFP1 183620
10q24.2	Leigh syndrome due to cytochrome c oxidase deficiency	256000	Mt, AR	3	COX3 603646
17p12	Leigh syndrome due to mitochondrial COX4 deficiency	256000	Mt, AR	3	COX4 602125

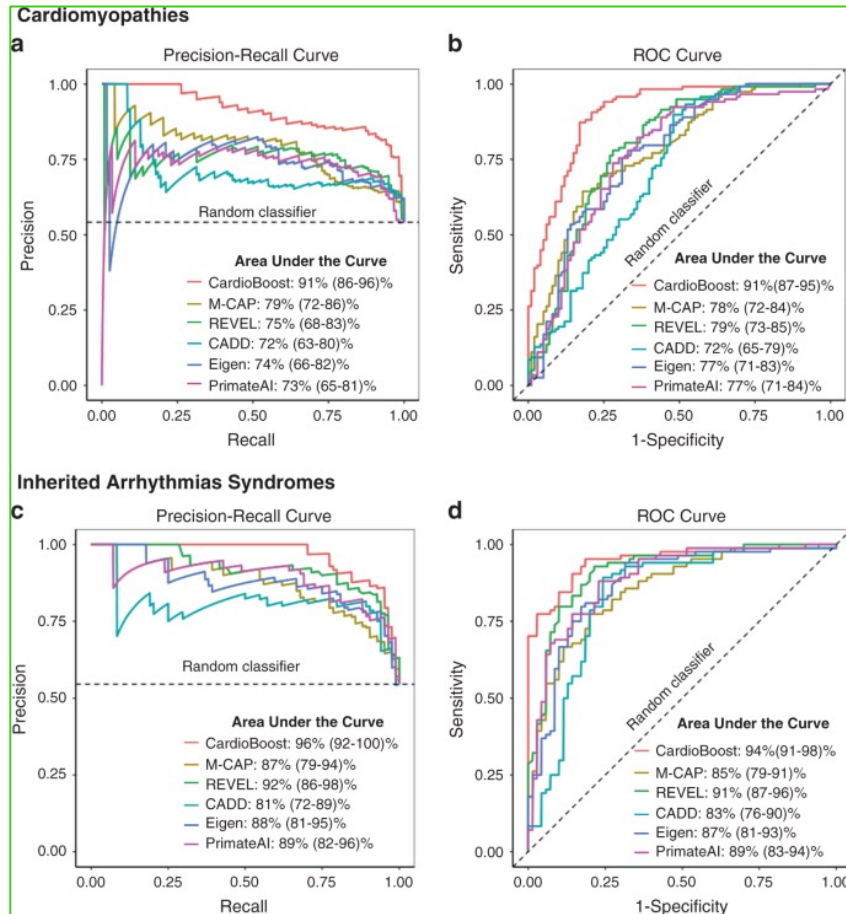
연관 유전자

INHERITANCE	증상
- Autosomal recessive	
- Mitochondrial	
GROWTH	
Other	
- Failure to thrive	
HEAD & NECK	
Eyes	
- Ophthalmoplegia	
- Optic atrophy	
- Nystagmus	
- Strabismus	
- Ptosis	
- Pigmentary retinopathy	
RESPIRATORY	
- Abnormal respiratory patterns	
- Respiratory failure	
SKIN, NAILS, & HAIR	
Hair	
- Hypertrichosis	
MUSCLE, SOFT TISSUES	
- Hypotonia	
NEUROLOGIC	
Central Nervous System	
- Psychomotor retardation	
- Hypotonia	
- Ataxia	
- Dystonia	
- Dysarthria	

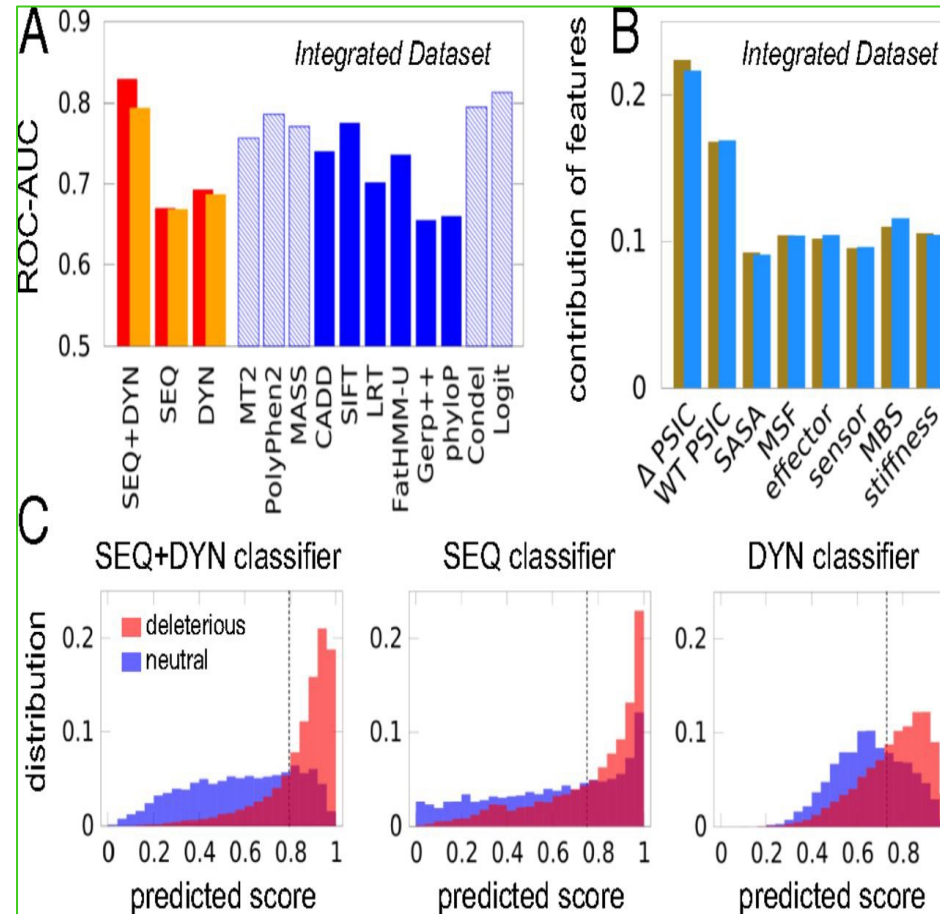
- Comparison with known disease gene - phenotype information
 - Largely subjective evaluation on phenotype similarity
 - Co-evaluation of phenotype and variant pathogenicity is also subjective in general.

Pathogenic Variant Prioritization

Variant pathogenicity prediction



(Zhang et al., Genetics in Medicine 2020)



(Ponzoni and Bahar, PNAS 2018)

- Ongoing development of variant pathogenicity prediction software using various characteristics
- NA/AA sequence characteristics of pathogenic variant
- Protein structure and function
- Ensemble integration of multiple prediction tools

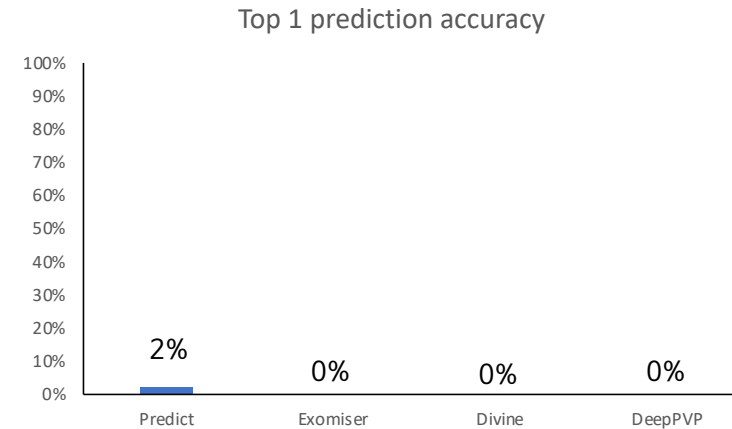
Pathogenic Variant Prioritization: Limitation

Incomplete coverage of genomic variation

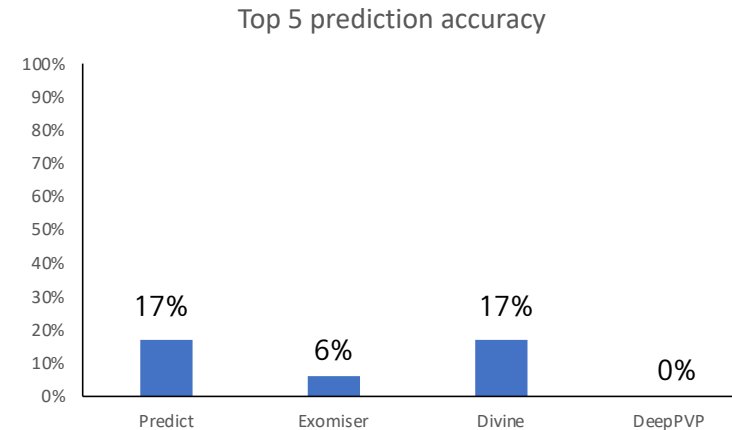
- **Most clinical applications target specific genomic regions**
 - Selected disease genes
 - Coding regions
- **Missing genomic regulations beyond DNA sequence**
 - Limited utilization of gene expression & protein information
- **Missing tissue-specificity**
 - Most clinical applications rely on germline DNA from blood cells

Low accuracy of pathogenic variant prioritization

- **Patients usually have multiple likely-pathogenic variants.**



* Evaluation using the data of 108 patients with confirmed diagnosis



Phenotype Matching

Curation of disease gene-phenotypes

The Human Gene Mutation Database

The Human Gene Mutation Database (HGMD) represents an attempt to collate all known (published) gene lesions responsible for human inherited disease and is maintained in Cardiff by D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, K. Evans, S. Heywood, M.J. Hayden, M.M. Chapman, M.E. Mort, L. Azevedo and D.S. Millar.

Table:

Table:	Description:
Gene symbol	The gene description, gene symbol (as recommended by the International Union of Pure and Applied Chemistry) and gene symbol has not yet been made official, a cDNA reference sequences are provided.
Genomic coordinates	Genomic (chromosomal) coordinates have been determined.
HGVSNomenclature	Standard HGVSNomenclature has been used.
Missense/nonsense	Single base-pair substitutions in coding region. Missense mutations are indicated by the letter 'M' and nonsense mutations by the letter 'N'.
Splicing	Mutations with consequences for mRNA splicing or acceptor splice site. Positions given as nucleotide positions relative to the start of the coding sequence.
Regulatory	Substitutions causing regulatory abnormalities such as transcriptional initiation site, initiation site, and transcription factor binding sites.
Small deletions	Micro-deletions (20 bp or less) are present. The first nucleotide of the deleted region is preceded in the given sequence by a hyphen and a number.
Small insertions	Micro-insertions (20 bp or less) are present. The first nucleotide of the inserted region is preceded in the given sequence by a plus sign and a number.
Small indels	Micro-indels (20 bp or less) are present. The first nucleotide of the indel region is preceded in the given sequence by a hyphen and a number.
Gross deletions	Information regarding the nature and location of gross deletions.
Gross insertions	Information regarding the nature and location of gross insertions.
Complex rearrangements	Information regarding the nature and location of complex rearrangements.
Repeat variations	Information regarding the nature and location of repeat variations.

OMIM®
Online Mendelian Inheritance in Man®
An Online Catalog of Human Genes and Genetic Disorders
Updated October 22, 2021

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map
Need help? : Example Searches, OMIM Search Help, OMIM Video Tutorials
Mirror site : <https://mirror.omim.org>

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

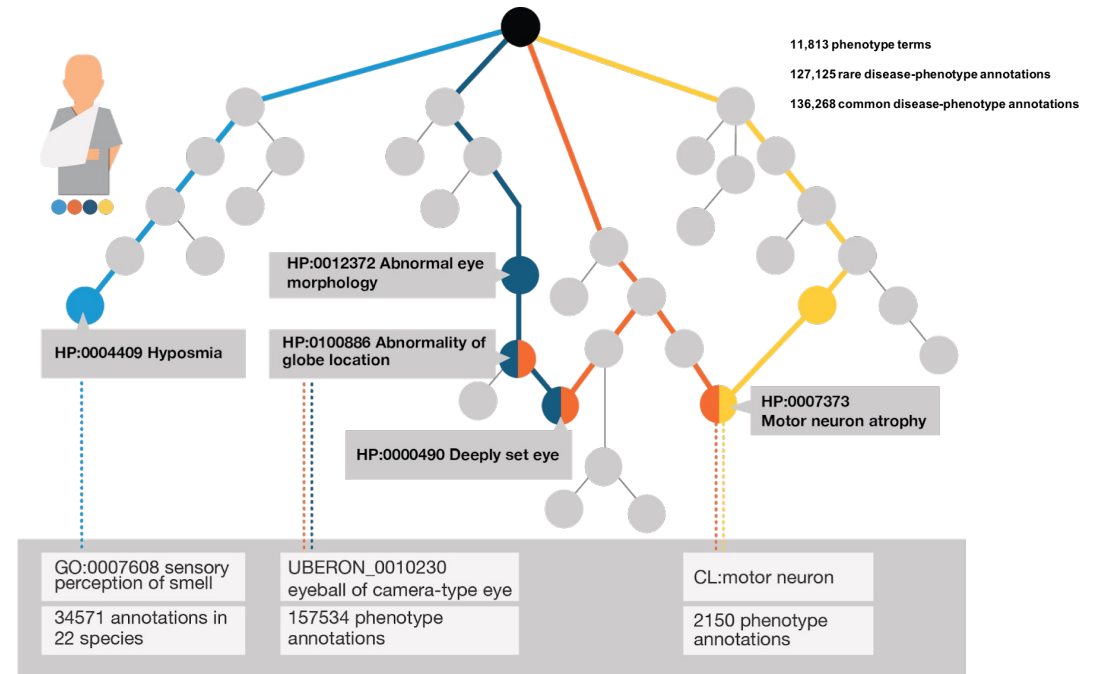
Make a donation!

DEANUSICK-NATHANS Department of Genetic Medicine
JOHNS HOPKINS MEDICINE

Follow us on Twitter

Online Mendelian Inheritance in Man

Standardization of phenotypes



- **Human Phenotype Ontology (by Monarch initiative)**
 - Consortium of EMBL-EBI, Jackson lab, etc.
 - Tree-structured definition of phenotype ontology
 - More than 13,000 phenotype terms
 - More than 156,000 annotations to hereditary disease

Phenotype Matching: Challenge

308350 ICD+

DEVELOPMENTAL AND EPILEPTIC ENCEPHALOPATHY 1; DEE1

INHERITANCE
- X-linked recessive

HEAD & NECK
Head
- Decreased head circumference

RESPIRATORY
- Dyspnea

ABDOMEN
Gastrointestinal
- Dysphagia

NEUROLOGIC
Central Nervous System
- Seizures, intractable
- Myoclonic seizures
- Hypsarrhythmia
- Arrest of psychomotor development after seizure onset
- Mental retardation
- Dystonia
- Status dystonicus
- Choreoathetosis
- Quadriplegic dyskinesia
- Axial hypotonia
- Hypertonia
- Hyperreflexia
- Spasticity
- Enlarged ventricles
- MRI shows T2-weighted signals in the basal ganglia

MISCELLANEOUS
- Onset of seizures in first months of life (usually 4 to 7 months)
- Dyskinesias occur in a subset of patients later than seizures (6 to 12 months)
- Males are most severely affected, but females can also be affected




MOLECULAR BASIS
- Caused by mutation in the X-linked aristaless-related homeobox gene (ARX, [300382.0001](#))

Contributors: Cassandra L. Kniffin - revised : 12/26/2007
Creation Date: John F. Jackson : 6/15/1995
Edit History: ckniffin : 04/01/2010


300055 ICD+

INTELLECTUAL DEVELOPMENTAL DISORDER, X-LINKED, SYNDROMIC 13; MRXS13

INHERITANCE
- X-linked recessive

HEAD & NECK
Head
- Microcephaly 
Face
- Micrognathia 
- Facial hypotonia
Ears
- Large ears
Mouth
- High-arched palate
- Sialorrhea
Teeth
- Bruxism
Neck
- Short neck 

GENITOURINARY
External Genitalia (Male)
- Macroorchidism (described in 1 family)

SKELETAL
Feet
- Pes cavus 

MUSCLE, SOFT TISSUES
- Distal atrophy of the legs

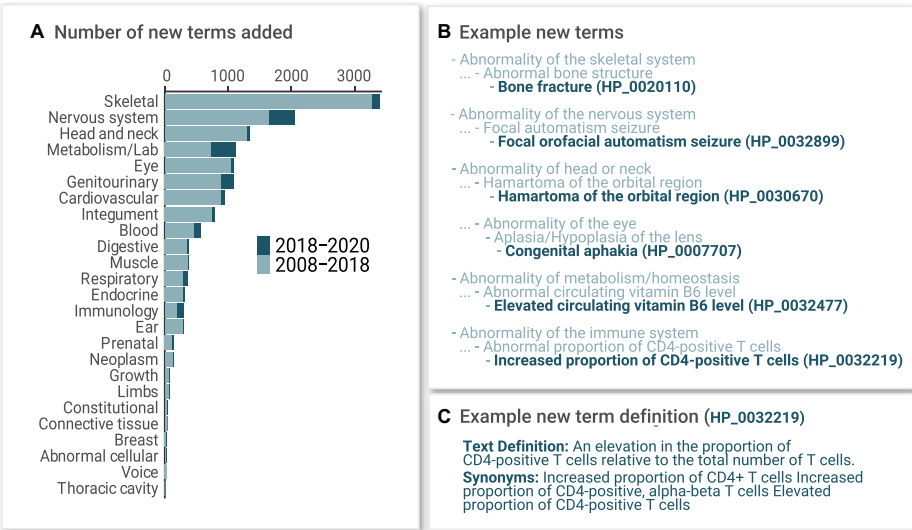
NEUROLOGIC
Central Nervous System
- Mental retardation
- Delayed development
- Delayed speech
- Spasticity
- Tremor
- Ataxia
- Parkinsonism
- Shuffling gait
- Spastic gait
- Hyperreflexia
- Increased tone

- A patient does not show all the previously reported phenotypes.
- Multiple diseases can show similar phenotypes.
- Matching known disease information with patient's phenotypes often requires expert clinician's involvement.

Advances in Utilizing Phenotype Information: HPO Example

Fine definition of phenotypes

“The Human Phenotype Ontology in 2021” (Kohler et al. NAR 2020)



Curating phenotype frequencies

- Vary rare (1 – 4%)
- Occasional (5 – 29%)
- Frequent (30 – 79%)
- Very frequent (80 – 99%)
- Obligate (100%)

Perrault Syndrome 3 OMIM:614129

Any Perrault syndrome in which the cause of the disease is a mutation in the CLPP gene.

Export Associations Report Entry Issue

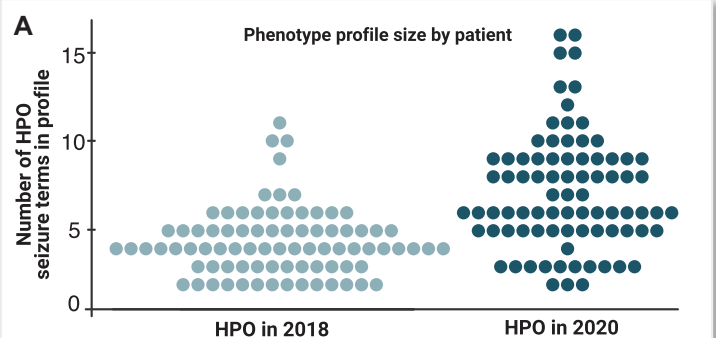
HPO Associations Gene Associations

Inheritance [1 annotation]

Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0000007	Autosomal recessive inheritance	-	-	OMIM

Growth [1 annotation]

Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0004322	Short stature	-	Occasional	OMIM



Developmental And Epileptic Encephalopathy 2 OMIM:300672

Any early infantile epileptic encephalopathy in which the cause of the disease is a mutation in the CDKLS gene.

Export Associations Report Entry Issue

HPO Associations Gene Associations

Inheritance [1 annotation]

Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0001423	X-linked dominant inheritance	-	-	OMIM

Digestive System [2 annotations]

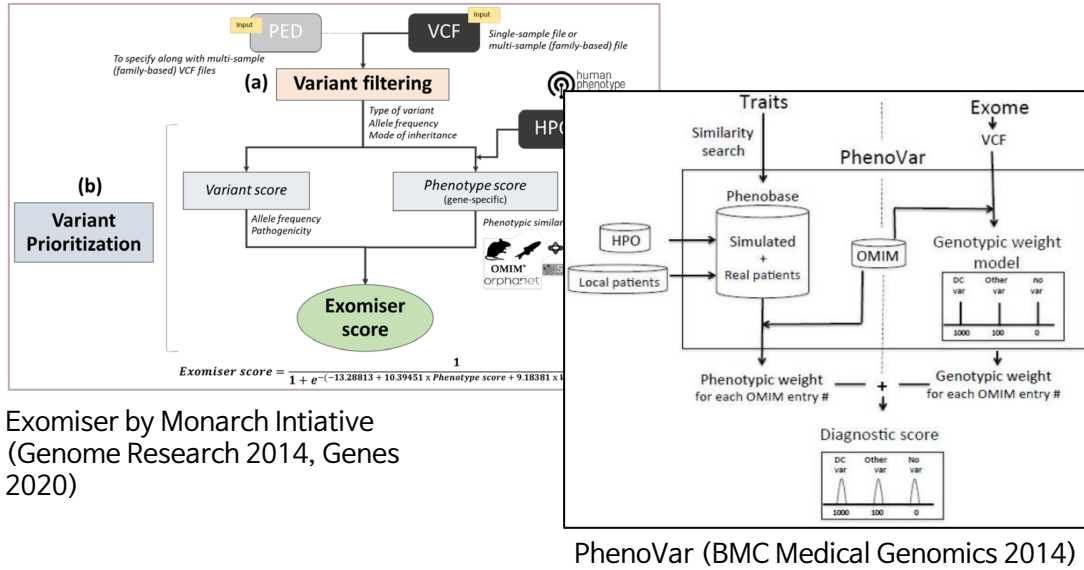
Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0002020	Gastroesophageal reflux	-	1/5	PubMed
HP:0002019	Constipation	-	3/5	PubMed

Skeletal system [1 annotation]

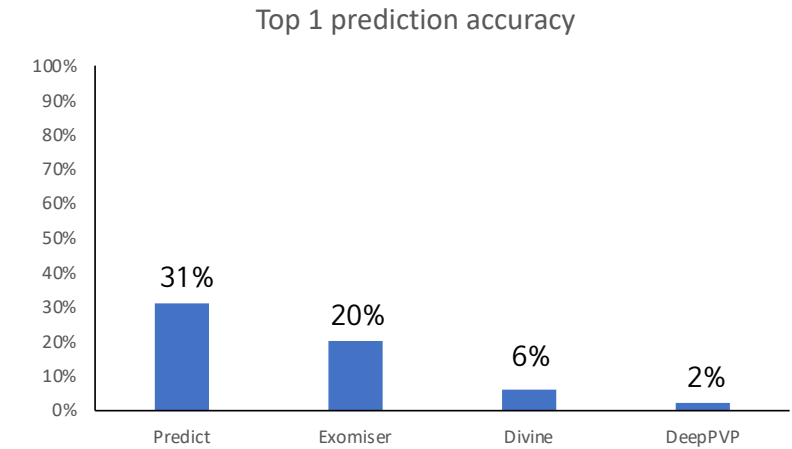
Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0002650	Scoliosis	-	4/5	PubMed

“Seizure” terms are increased from 68 to 348 by the seizure classification guideline from International League Against Epilepsy (ILAE).

Advances in Utilizing Phenotype Information: Integrated Tools



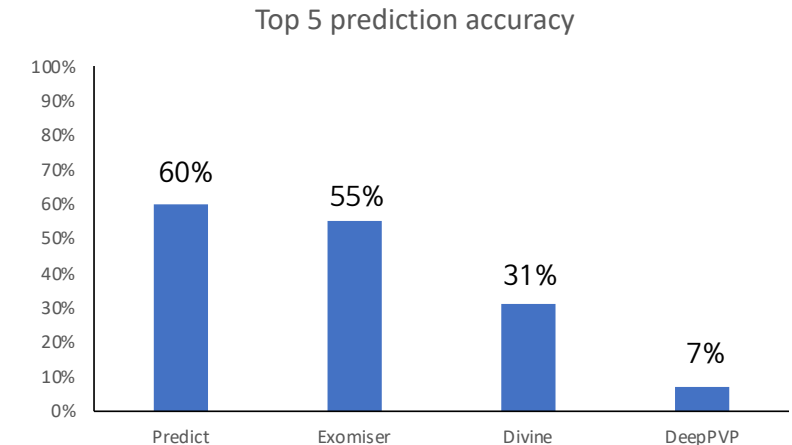
* Evaluation using the data of 108 patients with confirmed diagnosis



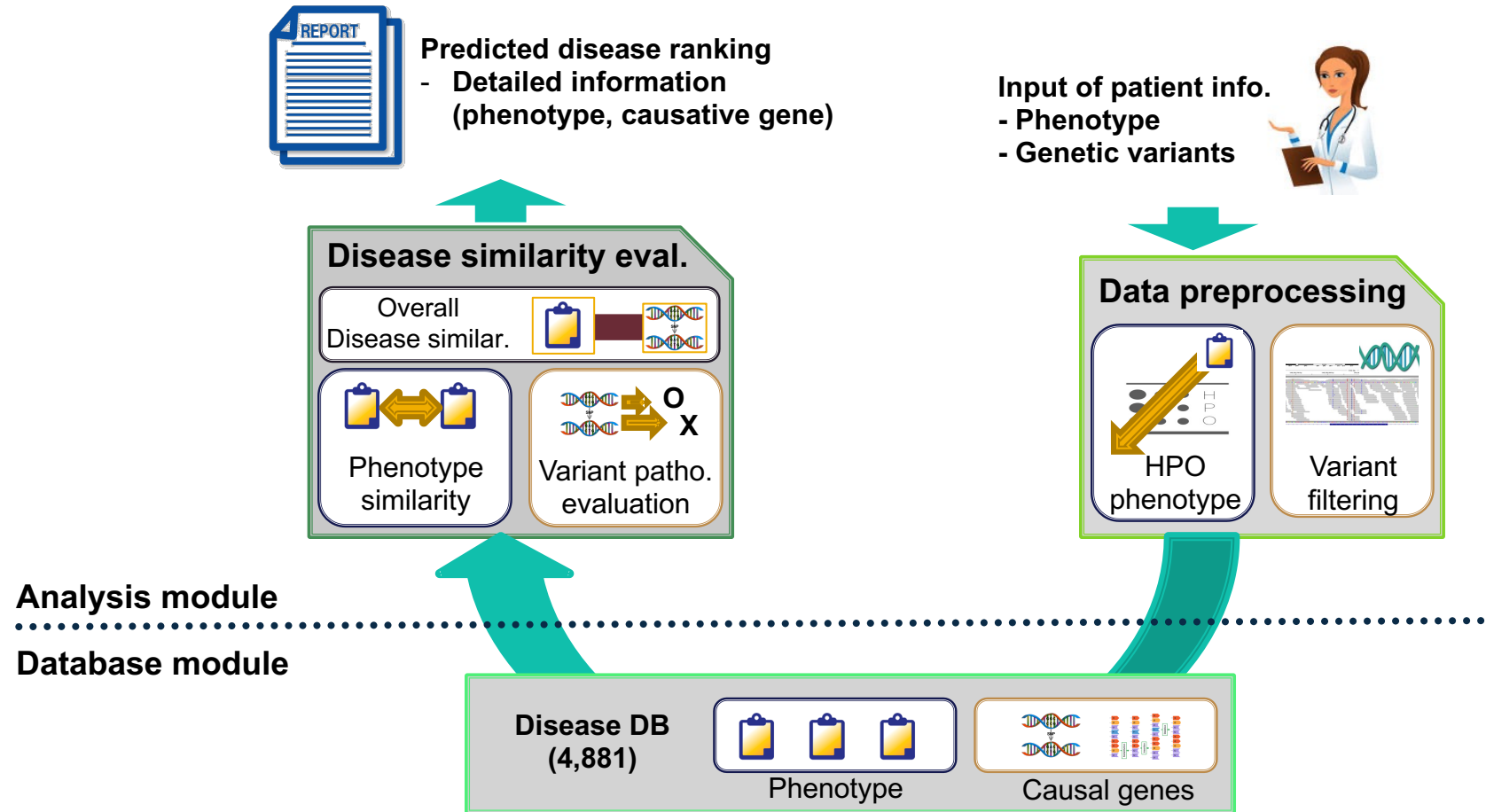
Divine (biorxiv 2018)

DeepPVP (BMC Bioinformatics 2019)

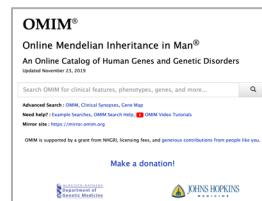
PREDICT (in preparation)



Rare Genetic Disorder Diagnosis System based on Data-Integrative Approach



Deciphering Developmental Disorders (DDD) project



OMIM



HPO

Key Components of Disease Likelihood Evaluation

Phenotype Similarity with Known Disease

환자의 표현형: Developmental regression, Seizure, Myoclonus, Respiratory failure, Brain atrophy

질환 A의 알려진 표현형: Global developmental delay, Diffuse cerebral atrophy, Microcephaly

- HPO-based phenotype similarity evaluation
- Optimizing evaluation using benchmark data

Variant Pathogenicity Evaluation

Input file: VCF, Annotation (Annovar-based annotation, Additional annotation with OMIM)

Filtering & prioritization: Frequency filtering, Functional impact filtering, Matching inheritance conditions & allelic status

Pathogenic prediction score: ClinVar, SIFT, LRT, PolyPhen2, MutationTaster, FATHMM, RadialSVM, LR

Pathogenic probability evaluation: Pathogenic probability prediction of genetic variant based on Bayesian model

Similarity evaluation: Predicting disease rank by calculating similarity between genetic variant and disease-related genetic information

Naive Bayes classifier: SIFT, Polyphen2, LRT, MutationTaster, FATHMM, RadialSVM, LR

Using 56,000 pathogenic ClinVar variants as training data

Variant pathogenicity-based ranking of candidate disease data

일렉 환자의 질환 유전자 후보	유전자 - 질환 연관 정보	순위 기반 유사도
유전자 A	A 연관 질환	0.9
유전자 B	B 연관 질환	0.7
유전자 C	C 연관 질환	0.2
변이별 pathogenicity	D 연관 질환	0

Evaluating the Final Suggestion Ranking of Disease

일렉 환자의 질환 유전자 후보	유전자 - 질환 연관 정보	순위 기반 유사도
유전자 A	A 연관 질환	0.9
유전자 B	B 연관 질환	0.7
유전자 C	C 연관 질환	0.2
변이별 pathogenicity	D 연관 질환	0

환자의 표현형	질환 1 표현형	순위 기반 유사도
Developmental regression	질환 1 표현형	0.41
Seizure	질환 2 표현형	0.57
Myoclonus	질환 3 표현형	0.34
Respiratory failure
Brain atrophy	질환 4,881 표현형	0.28

질환	Weighted sum of score
질환 1	0.18
질환 2	0.46
질환 3	0.79
...	...
질환 4,881	0.92

통합 데이터 기반 종합 질환 유사도

Evaluating disease likelihood based on variant pathogenicity and phenotype similarity

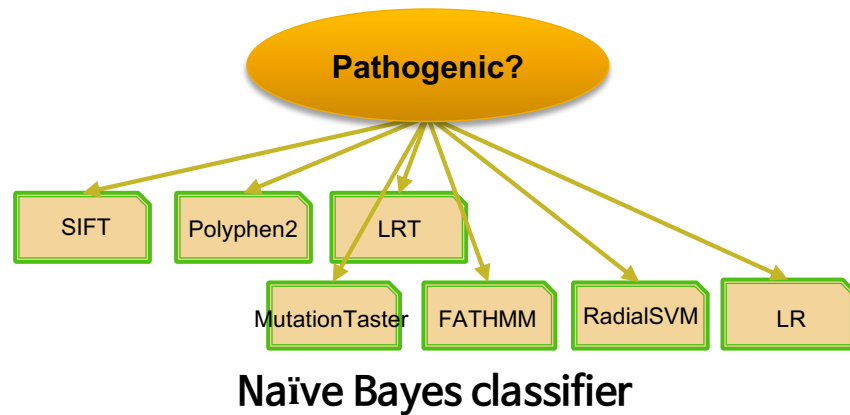
- Rep. of Korea Patent 10-2147847
- US Patent Application 16/879,584

$$Pr(D|P) = w_0 \times ecdf \left(D; \max_{v_i} \sum_{v_j} \frac{1}{T} P(v_i \text{ is pathogenic} | \text{PathoPred}_{v_j}) \right) + w_1 \times ecdf \left(D; \frac{1}{2} \frac{1}{m} \sum_{i=1}^m \max_{v_j} \left\{ \text{Resnick}(\text{phenotype}_{v_i}, \text{phenotype}_{v_j}) \times \min \left(\text{freq}_D(\text{phenotype}_{v_i}), \text{freq}_D(\text{phenotype}_{v_j}) \right) \right\} \right) + \frac{1}{n} \sum_{i=1}^n \max_{v_j} \left\{ \text{Resnick}(\text{phenotype}_{v_i}, \text{phenotype}_{v_j}) \times \min \left(\text{freq}_D(\text{phenotype}_{v_i}), \text{freq}_D(\text{phenotype}_{v_j}) \right) \right\} + w_2 \times ecdf \left(D; \sqrt{\frac{1}{T} \sum_{v_j} (\text{MRI}_j^p - \text{MRI}_j^d)^2} \right)$$

Variant Pathogenicity Prediction

Annotated pathogenic predictions from ANNOVAR

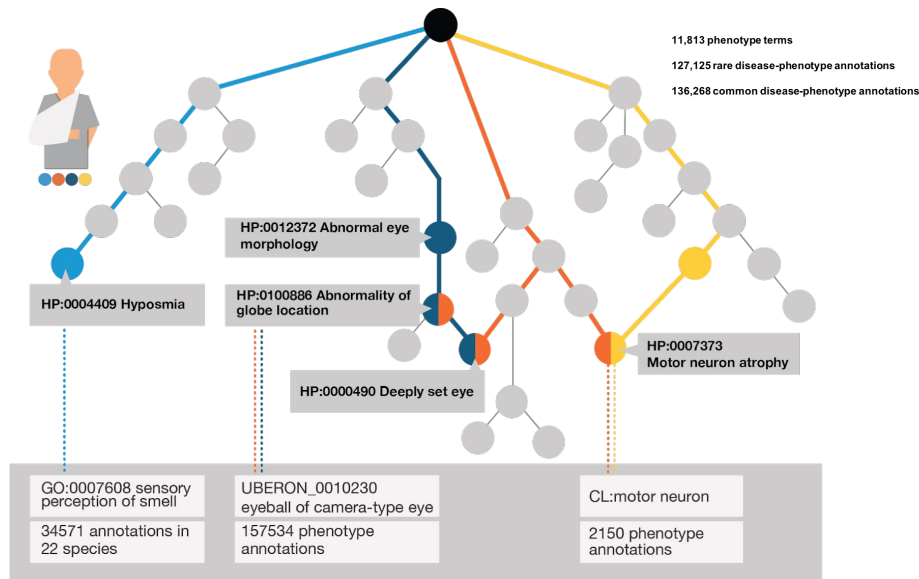
V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	R
CLINSIG	CLNDBN	CLNACC	CLNDSDB	CLNDSDB	SIFT_score	SIFT_prec	Polyphen2	Polyphen2	Polyphen2	Polyphen2	LRT_score	LRT_pred	Mutation	Mutation	Mutation	Mutation	FATHMM	FATHMM	R
Pathogenic	Immunodeficiency	RCV0001621	MedGen:OM	CN221808:6	0	D	0.352	N	1	D
Pathogenic	Immunodeficiency	RCV0001489	MedGen:OM	CN221808:6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pathogenic	Immunodeficiency	RCV0001489	MedGen:OM	CN221808:6	0	D	0.036	N	1	D
NA	NA	NA	NA	NA	0.05	D	1	D	0.999	D	0	U	1	D	0.975	L	-0.46	T	.
NA	NA	NA	NA	NA	0	D	1	D	0.999	D	0	U	0.999	D	0.975	L	1.18	T	.
NA	NA	NA	NA	NA	0.24	T	0.204	N	1	A



- Training data

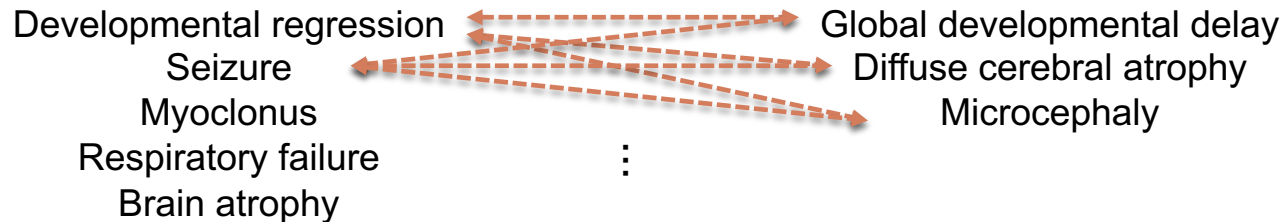
- 56,000 pathogenic, likely pathogenic variants from ClinVar
- Randomly selected 56,000 benign variants from normal subjects

Phenotype-based Similarity with Known Diseases



Patient's phenotype

Phenotype of Disease A



Ontology-based semantic similarity evaluation

Seven term-to-term similarity measures:

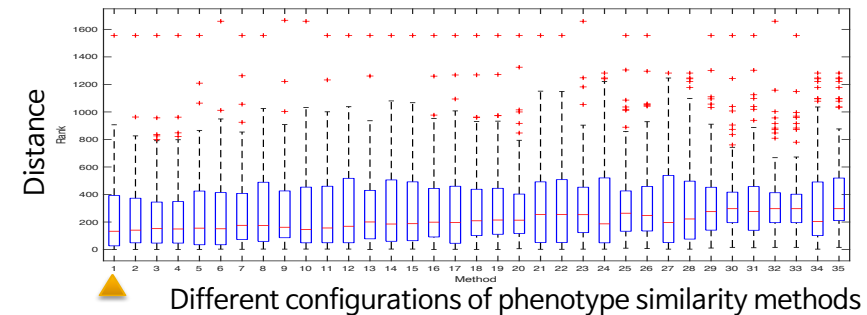
Information coefficient, Jiang-Conrath, Graph IC, Relevance, Wang, Lin, Resnik

Five term set similarity aggregation methods:

Max, Mean, funSimMax, funSimAvg, BMA

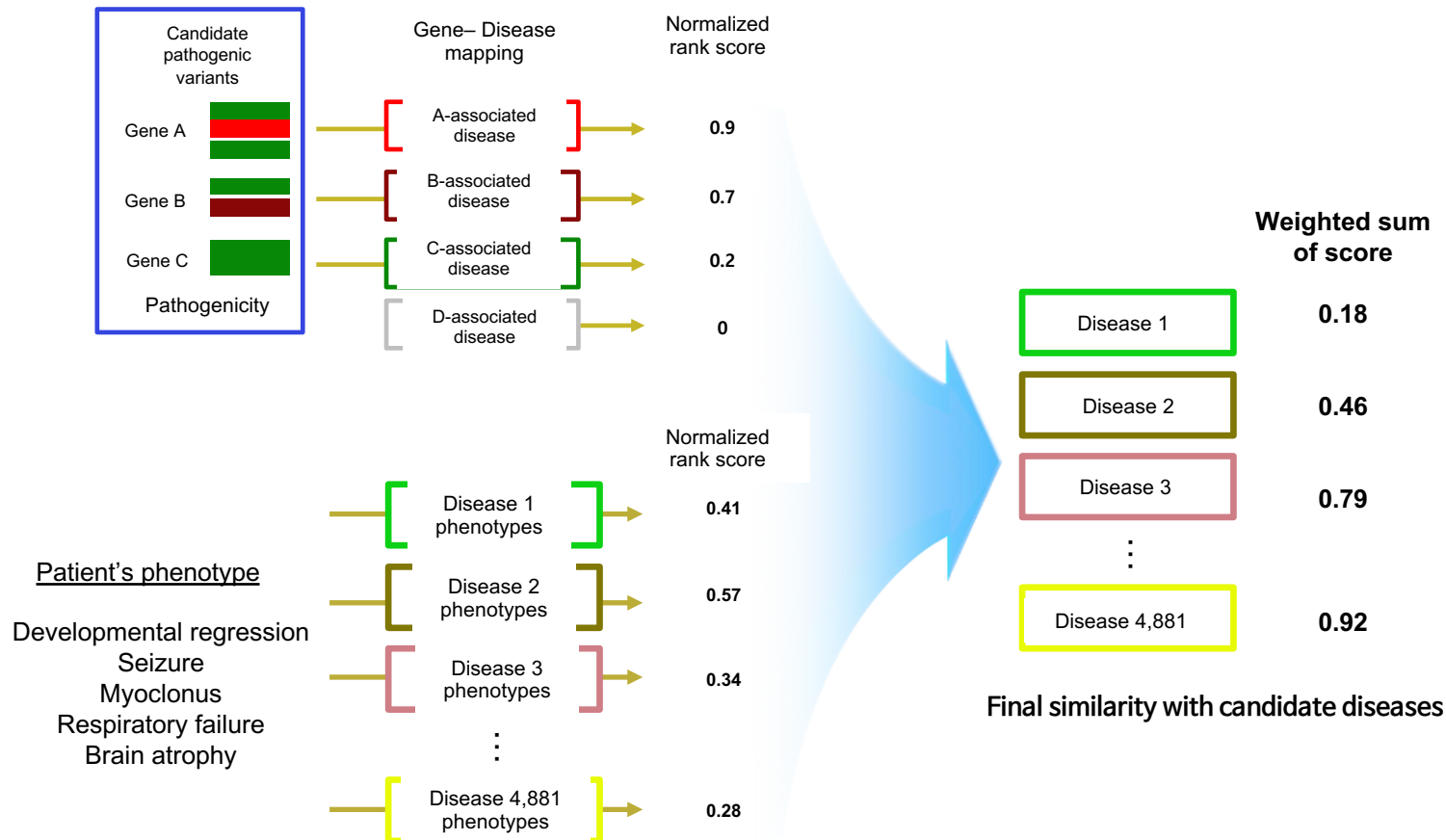
Identifying the optimal similarity evaluation method

- Using more than 100 patients' data as benchmark.



Final Evaluation of Disease Likelihood

Weighted sum of normalized disease rankings



- Evaluating disease likelihood based on variant pathogenicity and phenotype similarity

$$Pr(D|P) = w_0 \times ecdf \left(D; \max_{i,j} \frac{\sum_{t=1}^T \theta_t P(v_{ij} \text{ is pathogenic} | \text{PathoPred}_t)}{T} \right) + w_1 \times ecdf \left(D; \frac{1}{2} \left[\frac{1}{m} \sum_{i=1}^m \max_{1 \leq j \leq n} \left\{ \text{Resnick}(\text{phenotype}_{p_i}, \text{phenotype}_{D_j}) \times \min(\text{freq}_D(\text{phenotype}_{p_i}), \text{freq}_D(\text{phenotype}_{D_j})) \right\} \right] \right) + \frac{1}{n} \sum_{j=1}^n \max_{1 \leq i \leq m} \left\{ \text{Resnick}(\text{phenotype}_{p_i}, \text{phenotype}_{D_j}) \times \min(\text{freq}_D(\text{phenotype}_{p_i}), \text{freq}_D(\text{phenotype}_{D_j})) \right\} \right) + w_2 \times ecdf \left(D; \sqrt{\sum_f \gamma_f (\text{MRI}_f^p - \text{MRI}_f^D)^2} \right)$$

Benchmark Evaluation

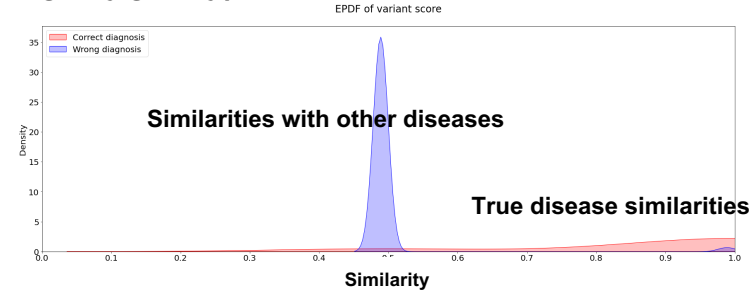
Test Cases with Known Diagnosis

번호	질환	번호	질환
1	(Epileptic encephalopathy)	56	epileptic encephalopathy, early infantile, 11
2	(Leigh Syndrome) cataracts, growth hormone deficiency, sensory neuropathy, sensorineural hearing loss, and skeletal dysplasia	57	Even-plus syndrome
3	(Leigh Syndrome) cataracts, growth hormone deficiency, sensory neuropathy, sensorineural hearing loss, and skeletal dysplasia	58	Faber lipogranulomatosis
4	(Leigh Syndrome) combined oxidative phosphorylation deficiency 13	59	GABA-transaminase deficiency
5	(Leigh Syndrome) Leigh syndrome due to mitochondrial complex I deficiency	60	GLUT1 deficiency syndrome
6	(Leigh Syndrome) Leigh syndrome due to mitochondrial complex I deficiency	61	GM1-gangliosidosis, type I
7	(Leigh Syndrome) Leigh syndrome, due to COX IV deficiency	62	GM1-gangliosidosis, type I
8	(Leigh Syndrome) mitochondrial complex I deficiency	63	Harel-Yoon syndrome
9	(Leigh Syndrome) Mitochondrial complex I deficiency	64	Hyperekplexia 3
10	(Leigh Syndrome) Mitochondrial short-chain enoyl-CoA hydratase 1 deficiency	65	infantile neuroaxonal dystrophy 1
11	(Leigh Syndrome) Thiamine metabolism dysfunction syndrome 2 (biotin- or thiamine-responsive encephalopathy type 2)	66	LCHAD deficiency
12	(Leigh Syndrome) Thiamine metabolism dysfunction syndrome 2 (biotin- or thiamine-responsive encephalopathy type 2)	67	Lesch-Nyhan syndrome
13	(Leigh Syndrome) Thiamine metabolism dysfunction syndrome 2 (biotin- or thiamine-responsive encephalopathy type 2)	68	(Kelley-Seegmiller syndrome)
14	(Rett syndrome like) epilepsy, focal with speech disorder and with or without mental retardation	69	Lethal congenital contracture syndrome 7 (LCCS7)
15	(Rett syndrome like) Epileptic encephalopathy, early infantile, 2	70	Leukodystrophy, hypomyelinating, 11
16	(Rett syndrome like) Epileptic encephalopathy, early infantile, 4	71	Leukodystrophy with vanishing white matter
17	(Rett syndrome like) Glass syndrome	72	Leukodystrophy, hypomyelinating, 6
18	(Rett syndrome like) mental retardation, autosomal dominant 19	73	Lubs X-linked mental retardation syndrome (MRXSL)
19	(Rett syndrome like) mental retardation, autosomal dominant 19	74	mental retardation, autosomal dominant 19
20	(Rett syndrome like) mental retardation, autosomal dominant 6	75	Mental retardation, autosomal dominant 35
21	(Rett syndrome like) mental retardation, autosomal dominant 6	76	Mental retardation, autosomal dominant, 9
22	(Rett syndrome like) myoclonic-atic epilepsy	77	mental retardation, X-linked
23	(Rett syndrome like) salt and pepper developmental regression syndrome	78	Mental retardation, X-linked, with cerebellar hypoplasia and distinctive facial appearance
24	Alexander disease	79	Mental retardation, X-linked, with cerebellar hypoplasia and distinctive facial appearance
25	alpha thalassemic with mental retardation syndrome	80	microcephaly 2, with or without cortical malformations
26	alpha thalassemic with mental retardation syndrome	81	Mowat-Wilson syndrome
27	Ataxia, early-onset, with oculomotor apraxia and hypalbuminemia	82	Muscular dystrophy-dystroglycanopathy, type C, 14
28	Bainbridge-Ropers syndrome	83	Muscular dystrophy-dystroglycanopathy, type C, 14
29	Bainbridge-Ropers syndrome	84	Nemaline myopathy 8
30	central core disease	85	Neurodevelopmental disorder with involuntary movement
31	Cerebellar ataxia, mental retardation, and dysequilibrium syndrome 2	86	Neurodevelopmental disorder with involuntary movement
32	ceroid lipofuscinosis, neuronal, 6	87	Neurodevelopmental disorder with or without hyperkinetic movements and seizures, autosomal dominant
33	Ceroid lipofuscinosis, neuronal, 6 (CLN6)	88	Ogden syndrome
34	Charcot-Marie-Tooth disease	89	Osteogenesis imperfecta, type I
35	Charcot-Marie-Tooth disease 4A	90	Pitt-Hopkins syndrome
36	Cockayne syndrome	91	Progressive myoclonic epilepsy
37	CODAS syndrome	92	Rett syndrome
38	Combined oxidative phosphorylation deficiency 13 (COXPD13)	93	Rigidity and multifocal seizure syndrome, lethal neonatal
39	combined oxidative phosphorylation deficiency 24	94	Schaaf-Yang syndrome
40	combined oxidative phosphorylation deficiency 24	95	Short stature, onychodysplasia, facial dysmorphism, and hypotrichosis
41	common variable immunodeficiency, type 10	96	Spastic ataxia, Charlevoix-Saguinay type
42	Congenital contractures of the limbs and face, hypotonia, and developmental delay	97	spastic paraplegia 3
43	Cornelia de Langer syndrome 1	98	spastic paraplegia 3
44	Cutis laxa, autosomal recessive type IIa	99	spastic paraplegia 3
45	D-bifunctional protein deficiency	100	Spastic paraplegia 43
46	Dravet syndrome	101	spastic paraplegia 5A
47	Dravet syndrome	102	spastic paraplegia 8
48	Dystonia 24	103	Spinal muscular atrophy, distal, autosomal recessive 1 (DSMA1)
49	Encephalopathy, acute, infection-induced, susceptibility to, 3 (IAE3)	104	spinal muscular atrophy, lower extremity-predominant 1
50	Epilepsy, pyridoxine-dependent	105	Spinocerebellar ataxia, 13
51	Epileptic encephalopathy	106	Spinocerebellar ataxia, 15
52	Epileptic encephalopathy, early infantile, 2	107	Spinocerebellar ataxia, autosomal recessive 1 (SCAR1)
53	Epileptic encephalopathy, early infantile, 31	108	Waardenburg syndrome
54	Epileptic encephalopathy, early infantile, 31 (EIEE31)	109	Wiesacker-Wolff syndrome
55	Epileptic encephalopathy, early infantile, 45	110	Yunis-Varon syndrome

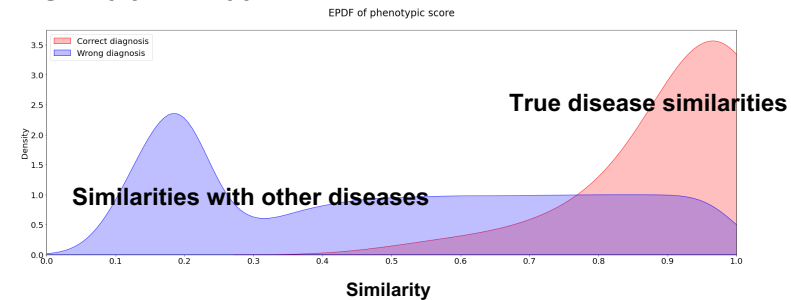
110 cases
(93 different diseases)

Distribution of Similarity with True Disease

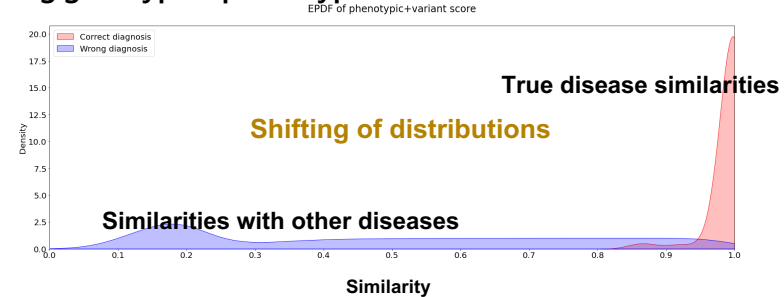
Using only genotype information



Using only phenotype information



Using genotype + phenotype



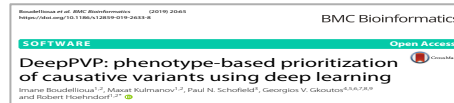
Comparative Benchmark Evaluation

Comparative evaluation with other software tools

번호	질환	번호	질환
1	(Epileptic encephalopathy)	56	epileptic encephalopathy, early infantile, 11
2	(Leigh Syndrome) cataracts, growth hormone deficiency, sensory neuropathy, sensorineural hearing loss, and skeletal dysplasia	57	Even-plus syndrome
3	(Leigh Syndrome) cataracts, growth hormone deficiency, sensory neuropathy, sensorineural hearing loss, and skeletal dysplasia	58	Farber lipogranulomatosis
4	(Leigh Syndrome) combined oxidative phosphorylation deficiency 13	59	GABA-transaminase deficiency
5	(Leigh Syndrome) Leigh syndrome due to mitochondrial complex II deficiency	60	GLUT1 deficiency syndrome
6	(Leigh Syndrome) Leigh syndrome due to mitochondrial complex II deficiency	61	GLUT1 deficiency syndrome
7	(Leigh Syndrome) Leigh syndrome, due to COX IV deficiency	62	GM1-gangliosidosis, type I
8	(Leigh Syndrome)	63	GM1-gangliosidosis, type I
9	(Leigh Syndrome) M	64	Harel-Yoon syndrome
10	(Leigh Syndrome) M	65	Hyperekplexia 3
11	(Leigh Syndrome) M	66	infantile neuroaxonal dystrophy 1
12	(Leigh Syndrome) Thiamine-responsive encephalopathy type 2	67	LCHAD deficiency
13	(Leigh Syndrome) Thiamine metabolism dysfunction syndrome 2		
14	(Rett syndrome like) epilepsy, focal with speech disorder and with or without mental retardation	73	Lubs X-linked mental retardation syndrome (MRXSL)
15	(Rett syndrome like) Epileptic encephalopathy, early infantile, 2	74	mental retardation, autosomal dominant 19
16	(Rett syndrome like) Epileptic encephalopathy, early infantile, 4	75	Mental retardation, autosomal dominant 35
17	(Rett syndrome like) Glass syndrome	76	Mental retardation, autosomal dominant, 9
18	(Rett syndrome like) mental retardation, autosomal dominant 19	77	mental retardation, X-linked
19	(Rett syndrome like) mental retardation, autosomal dominant 19	78	Mental retardation, X-linked, with cerebellar hypoplasia and distinctive facial appearance
20	(Rett syndrome like) mental retardation, autosomal dominant 6	79	Mental retardation, X-linked, with cerebellar hypoplasia and distinctive facial appearance
21	(Rett syndrome like) mental retardation, autosomal dominant 6	80	microcephaly 2, with or without cortical malformations
22	(Rett syndrome like) myoclonic-atonic epilepsy	81	Mowat-Wilson syndrome
23	(Rett syndrome like) salt and pepper developmental regression syndrome	82	Muscular dystrophy-dystroglycanopathy, type C, 14
		83	Muscular dystrophy-dystroglycanopathy, type C, 14
		84	Nemaline myopathy 8

110 test cases
(93 different diseases)

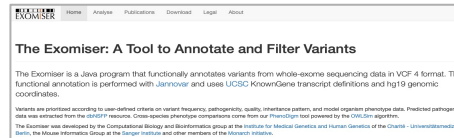
Other software tools



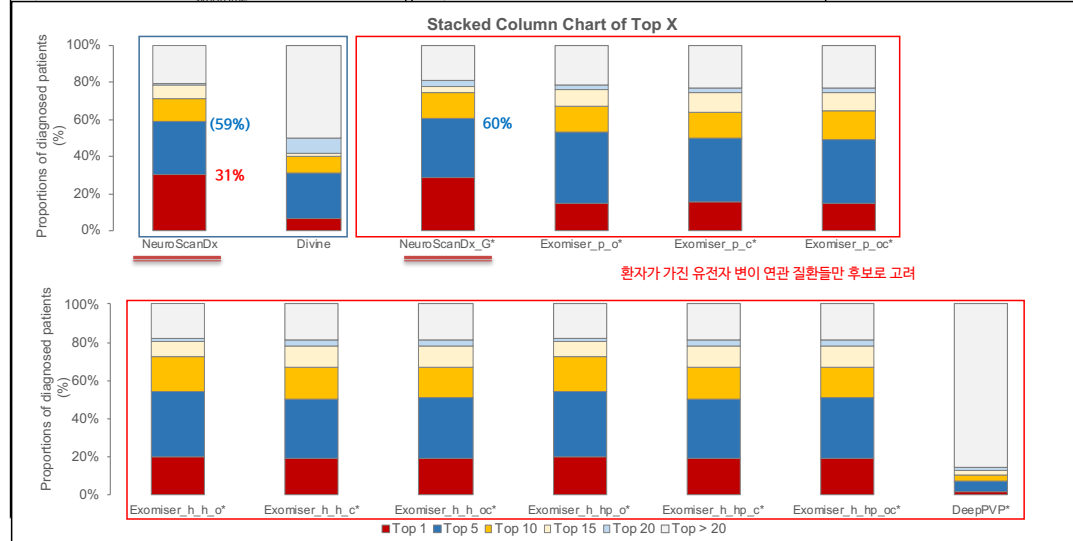
DeepPVP (2019)
By KAUST



Divine (2018)
By Cleveland Clinic, USA



Exomiser (2014, 2020)
By Monarch Initiative



Top 1 prediction accuracy

- Predict: 31%
- DeepPVP(2%), Divine(6%), Exomiser(20%)

Top 5 prediction accuracy

- Predict: 60%
- DeepPVP(7%), Divine(31%), Exomiser(55%)

PREDICT Web

(<http://sysbio.gachon.ac.kr/predict>)

PREDICT Home Search Disease About us

Welcome

BETA ***PREDICT : Prioritizing rare genetic disorders with combination of genotype and phenotypes***

A web-based software platform for prioritizing diseases of the patients using VCF and phenotypic data encoded using HPO (The Human Phenotype Ontology)

Email*

Result Access Password*

FULL institution name*

Genotype data* no file selected

Observed symptoms*

HPO ID	Features	Delete
...

Suspected disease

PREDICT Home Search Disease About us

Searching candidate disease

Job ID : 17, received at : 2023-05-22 16:36:06

Email : sjung@gachon.ac.kr

Institution : Gachon University

VCF file : P_NS_23.vcf

Observed symptoms : Seizures

Suspected disease : NULL

Search condition

*TOP 10

uses

- Phenotype
- VCF

Evidence List

Rank	Candidate gene	Disease name	OMIM ID	Detailed results
1	SYNGAP1	MENTAL RETARDATION, AUTOSOMAL DOMINANT 5; MRDS	612621	View
2	KIF5C	CORTICAL DYSPLASIA, COMPLEX, WITH OTHER BRAIN MALFORMATIONS 2; CDCBM2	615282	View
3	CDKL5	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 2; EIEE2	300672	View
4	ABCD1	ADRENOLEUKODYSTROPHY; ALD	300100	View
5	KIF1B	CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2A1; CMT2A1	118210	View
6	KIF1B	NEUROBLASTOMA, SUSCEPTIBILITY TO, 1; NBLST1	256700	View
7	TTN	TIBIAL MUSCULAR DYSTROPHY, TARDIVE; TMD	600334	View
8	NEB	NEMALINE MYOPATHY 2; NEM2	256030	View
9	TTN	SALIH MYOPATHY; SALMY	611705	View
10	TTN	MYOPATHY, MYOFIBRILLAR, 9, WITH EARLY RESPIRATORY FAILURE; MFM9	603689	View

1290 Seizures

2463	Language impairment
2353	EEG abnormality
1263	Global developmental delay
1290	Seizures
1270	Motor delay
200134	Epileptic encephalopathy
1290	Generalized hypotonia
252	Microcephaly
473	Torticollis
1249	Intellectual disability
2376	Developmental regression
729	Autistic behavior

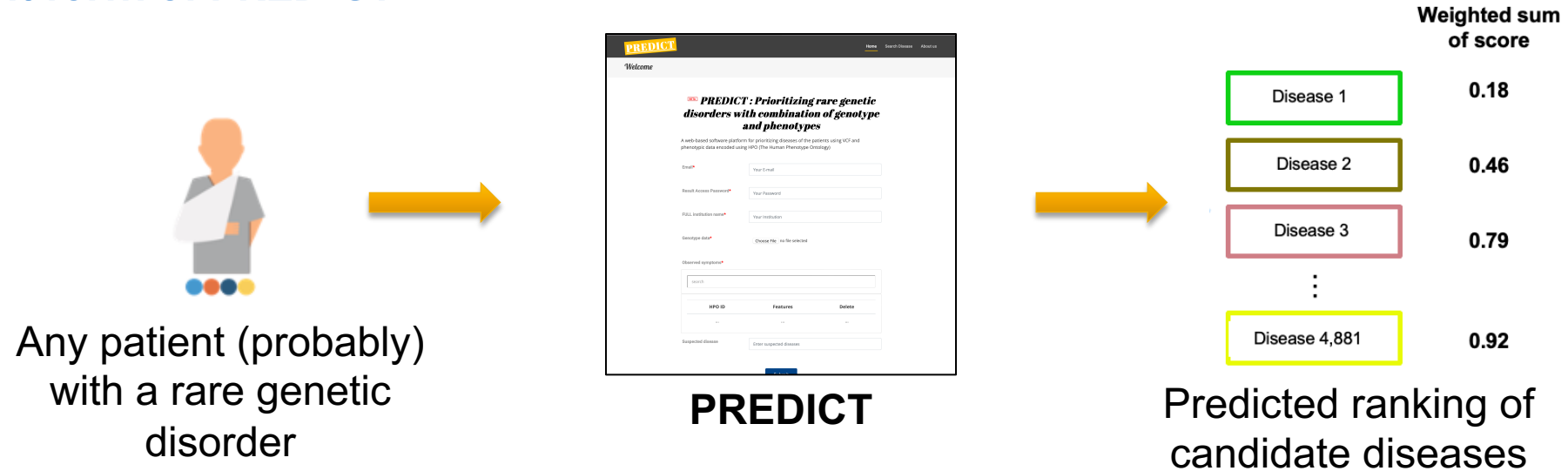
Evidence of genomic data

Evidence's variants

Gene	Disease	OMIM	Inheritance
SYNGAP1_MRDS	MENTAL RETARDATION, AUTOSOMAL DOMINANT 5; MRDS	612621	AD

Revision of PREDICT for Medical Application

◆ Current form of PREDICT



■ Challenges:

A clinical trial cannot cover all rare genetic disorders. - More focused target disease should be set.

How to prove clinical benefits - No previously approved similar devices, thus no reference. We should set our own rules.

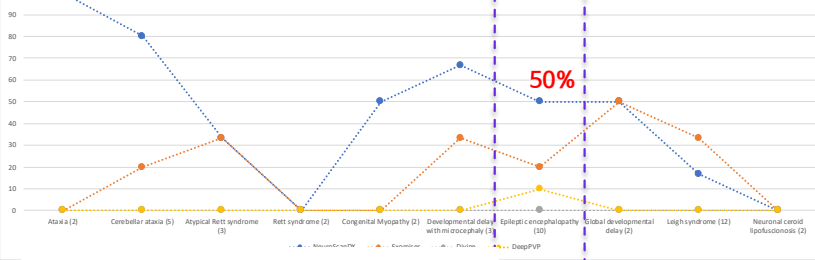
Revision of PREDICT for Medical Application

Set Application to Specific Disease

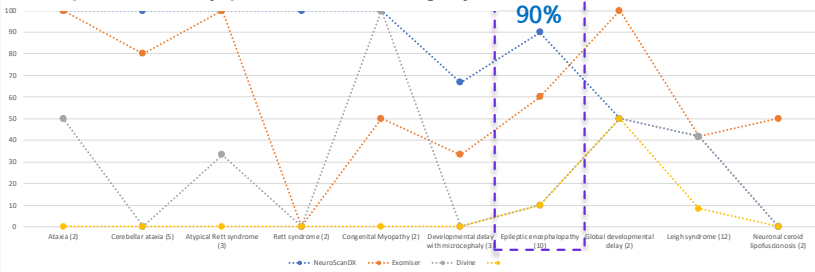
Target disease for medical device

- Cannot cover all known genetic disorders in clinical trials
- One disease category per one clinical trial

Top 1 accuracy per disease category



Top 5 accuracy per disease category

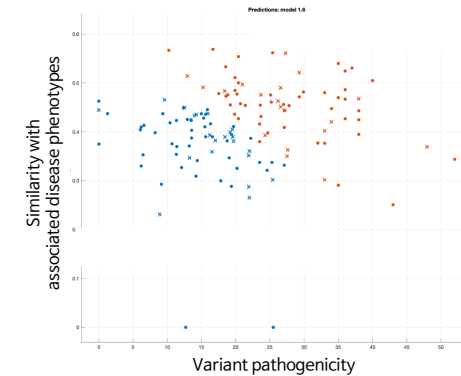


- One disease category is chosen for med. application.
- PREDICT can be optimized for specific disease.



1. Predicting if a patient has specific disease

- YES
- UNKNOWN



2. Predicting causative gene

예측 순위	병인 유전자	관련 질환
1	KMT2C	KLEEFSTRA SYNDROME 2; KLEFS2
2	FRG1	FACIOSCAPULOHUMERAL MUSCULAR DYSTROPHY 1; FSH
3	ADNP	HELMSMOORTEL-VAN DER AA SYNDROME; HVDAS
4	FRG1	INTESTINAL PSEUDOObSTRUCTION, NEURONAL, CHRONI
5	GNE	SIALURIA

Summary

- ◆ **Systematic approach to aid the diagnosis of rare genetic disorders**
 - Requires large patient cohorts for proper data curation
 - Requires various pattern recognitions and information processing
 - Well-designed systems can help clinicians to some extent.
 - Quick analysis for early diagnosis materials.
 - Provides bottom line diagnosis performance.
- ◆ **Ongoing challenges toward clinical applications**