

Informatics-Driven Efforts for the Diagnosis of Rare Genetic Disorders

BIOINFO 2021 OnBIT Award Lecture

Sungwon Jung, Ph.D

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

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

What is Rare Genetic Disorder?

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.



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 앉혀주면 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

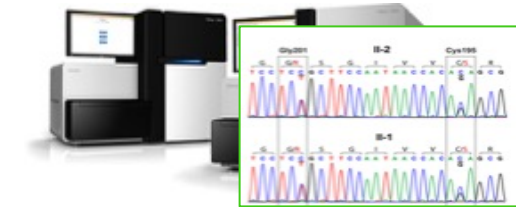
6월부터 조금씩 일어서기 시작

10월 현재 화장실 걸어서 가고 계단 오르기 시작

자료: 서울대병원 소아신경과



Clinical observation of patient's abnormal phenotypes



Identifying patient's genetic variations



Listing candidate diseases



Gene X



Previously reported > 6,000 disorders

Phenotypes
Causal genes

Prioritizing potentially causative pathogenic genetic variants

Final diagnosis of patient

Successful diagnosis is significantly dependent on clinical and technical abilities.

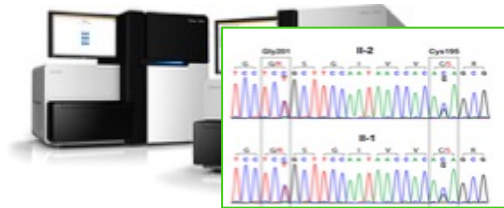
Diagnosis of Rare Genetic Disorders

Step 1: Clinical diagnosis

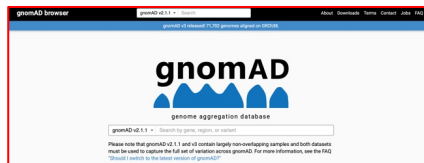


- Identifying abnormal phenotypes
- (Listing candidate diseases)

Step 2: Identifying rare genetic variants



Identifying genetic variations



Filtering genetic variants with low VAF

Step 3: Prioritizing pathogenic genetic variants

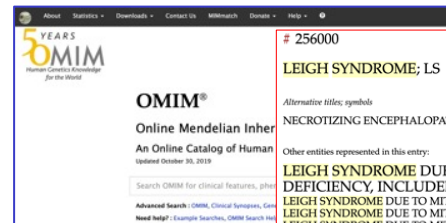
Possible large variation in diagnosis

	Strong		Pathogenic		
	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder (RA 1:85) 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/figure product BP4 1 substitution in gene/missense only nonconservative change BP3 Silent variant with non-predicted splice impact BP7 In-frame indels in repeat/short tandem repeats BP5	Multiple lines of computational evidence suggest a deleterious effect on the gene/figure product P03	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen P05 Frameshift 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 P06
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 & majority confirmed) P02	De novo (paternity and majority confirmed) P02	
Allelic data	Observed in homo with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		De novo (without parental & majority confirmed) P02	De novo (paternity and majority confirmed) P02	
Other databases	Reputable source critical shared data - benign BP6	Reputable source - pathogenic P05			
Other data	Found in case with an alternate cause BP5	Patient's phenotype or P1 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



256000
LEIGH SYNDROME; LS 질환명
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	ML, AR	3	BCSL1 603647
5p15.33	Leigh syndrome	256000	ML, AR	3	SDHA 603857
9q34.2	Leigh syndrome, due to COX IV deficiency	256000	ML, AR	3	SLRFP1 183620
10q24.2	Leigh syndrome due to cytochrome c oxidase deficiency	256000	ML, AR	3	COX3 603646
17p12	Leigh syndrome due to mitochondrial COX4 deficiency	256000	ML, AR	3	COX4 603125

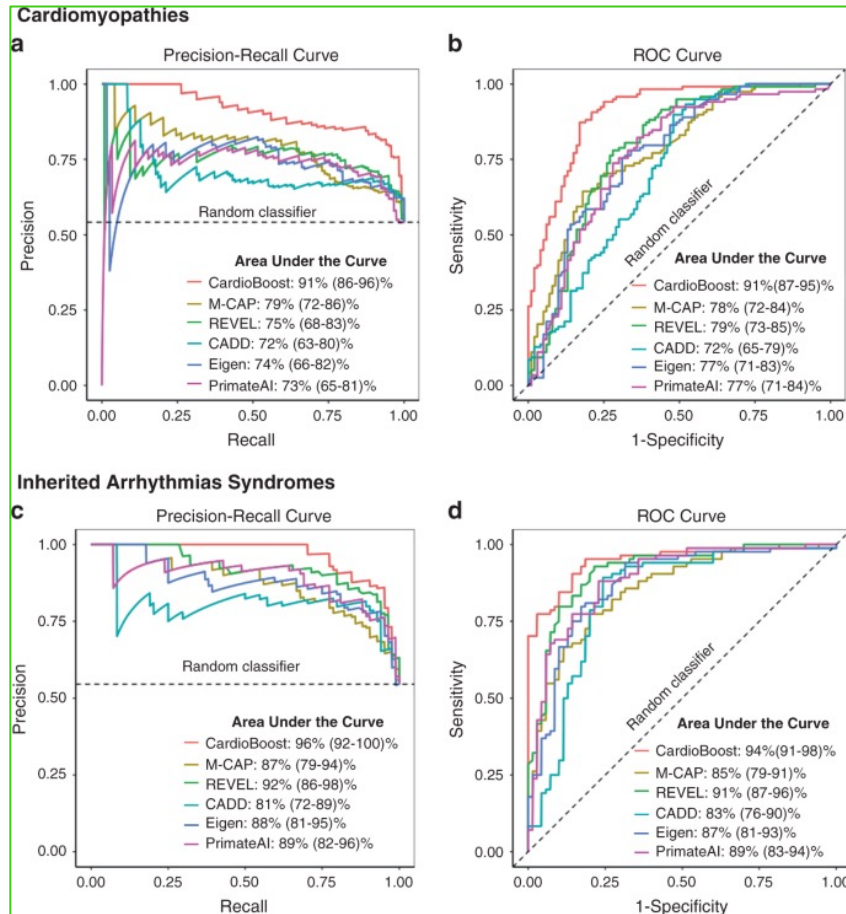
연관 유전자

INHERITANCE
- Autosomal recessive
- Mitochondrial
GROWTH
Other
- Failure to thrive
HEAD & NECK 증상
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

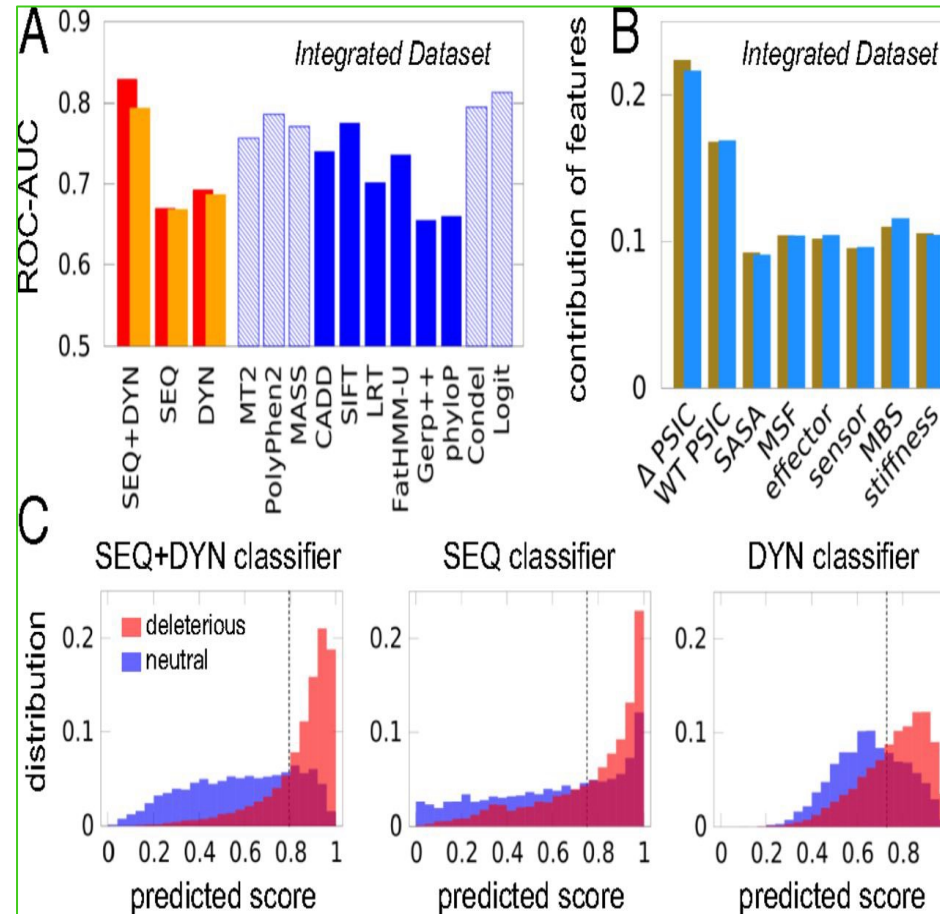
- Comparison with known disease gene - phenotype information
 - Largely subjective evaluation on phenotype similarity
 - Co-evaluation of phenotype and variant pathogenicity is also subjective in general.

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

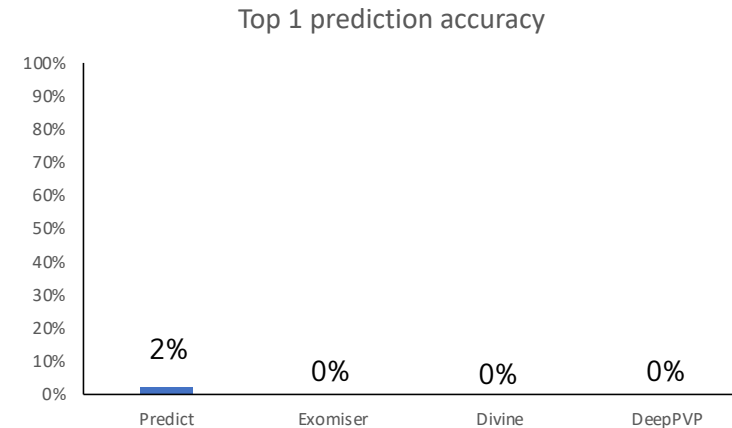
Pathogenic Variant Prioritization: Limitation

Incomplete coverage of genomic variation

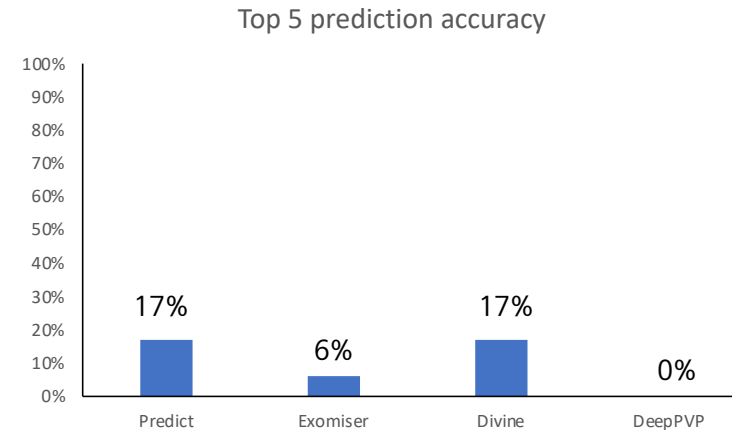
- **Most clinical applications target specific genomic regions**
 - Selected disease genes
 - Coding regions
- **Missing genomic regulations beyond DNA sequence**
 - Limited utilization of gene expression & protein information
- **Missing tissue-specificity**
 - Most clinical applications rely on germline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

Low accuracy of pathogenic variant prioritization

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



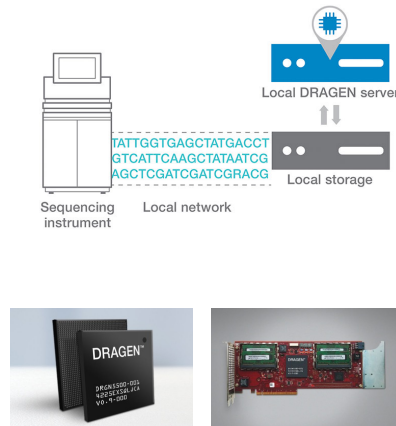
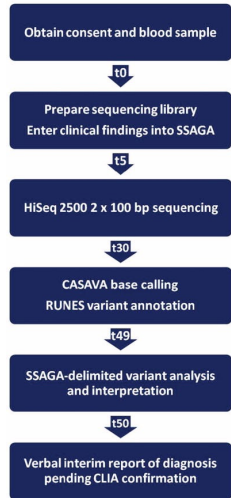
* Evaluation using the data of 108 patients with confirmed diagnosis (unpublished)



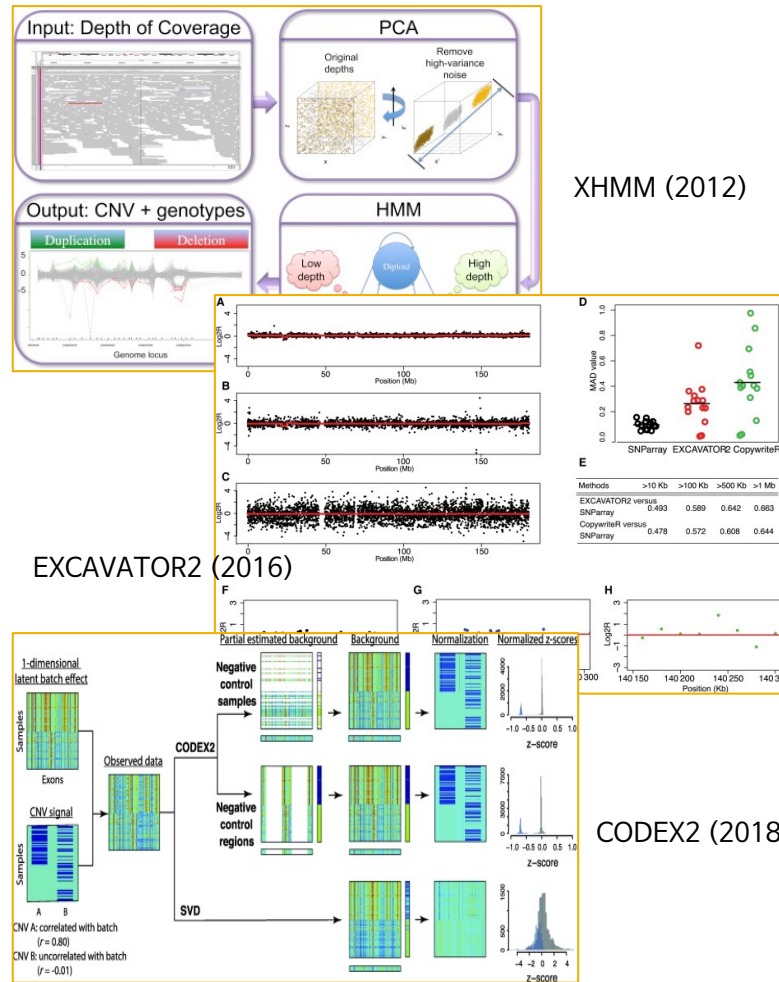
Advances in Pathogenic Variant Identification: Extending Coverage

Whole-genome sequencing

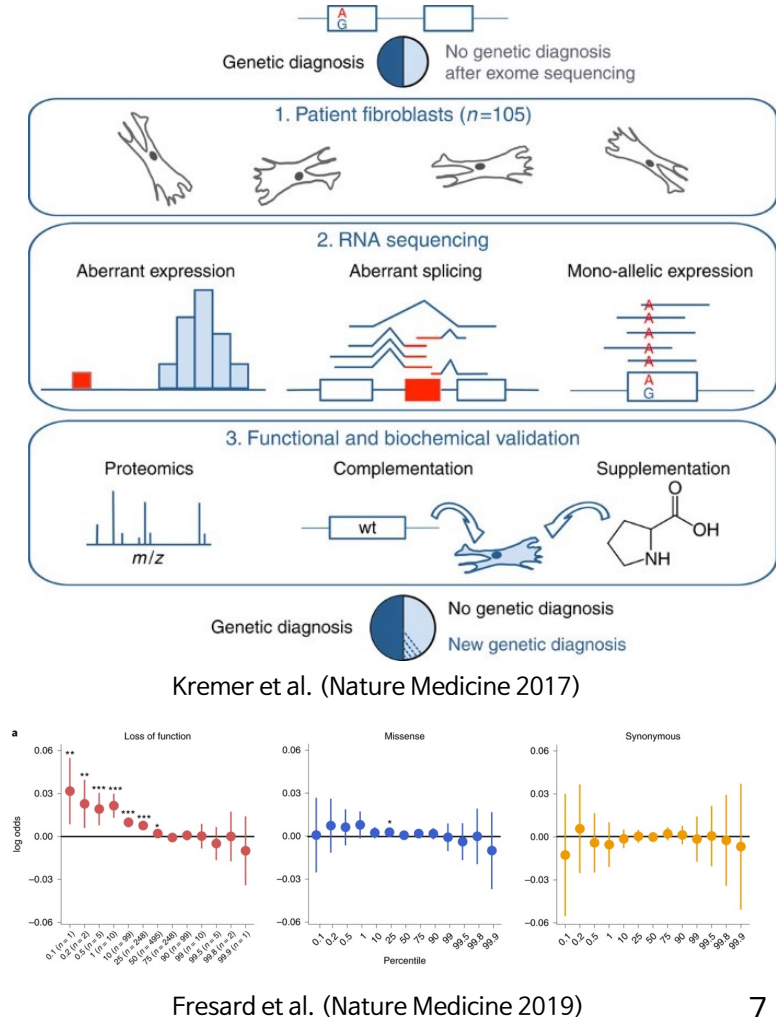
- Identifying non-coding/structural variants
- Rapid WGS such as STAT-Seq
 - 30~40X WGS variant analysis within 50 hrs
- Custom HW such as Illumina DRAGEN
 - 30X WGS analysis within 25 mins



SV identification with Targeted-seq



Application of RNA-seq



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 provided.
HGVSNomenclature	Standard HGVSNomenclature has been provided.
Missense/nonsense	Single base-pair substitutions in coding region, missense mutations, and nonsense mutations (premature stop codons) are listed.
Splicing	Mutations with consequences for mRNA splicing, such as splice site mutations, are listed.
Regulatory	Substitutions causing regulatory abnormalities, such as transcription start site mutations, are listed.
Small deletions	Micro-deletions (20 bp or less) are present and the first numbered codon is preceded in the given sequence.
Small insertions	Micro-insertions (20 bp or less) are present and the first numbered codon is preceded in the given sequence.
Small indels	Micro-indels (20 bp or less) are present and the first numbered codon is preceded in the given sequence.
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.

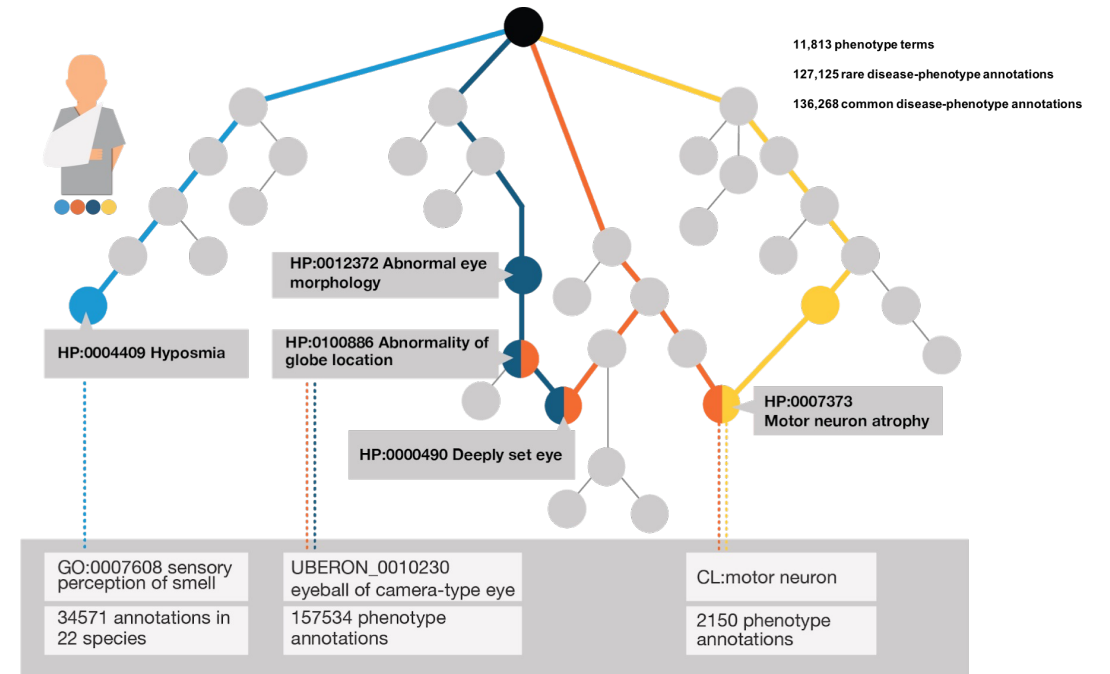
Make a donation!

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

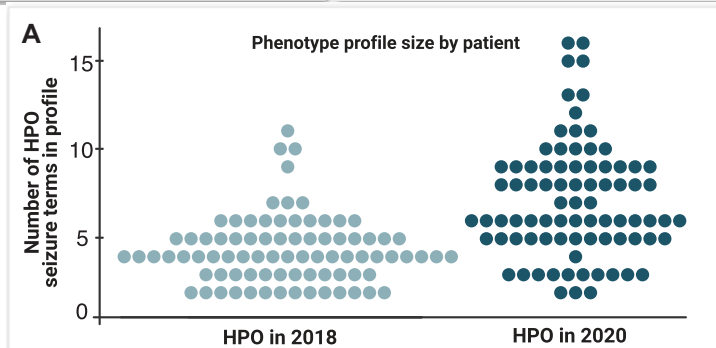
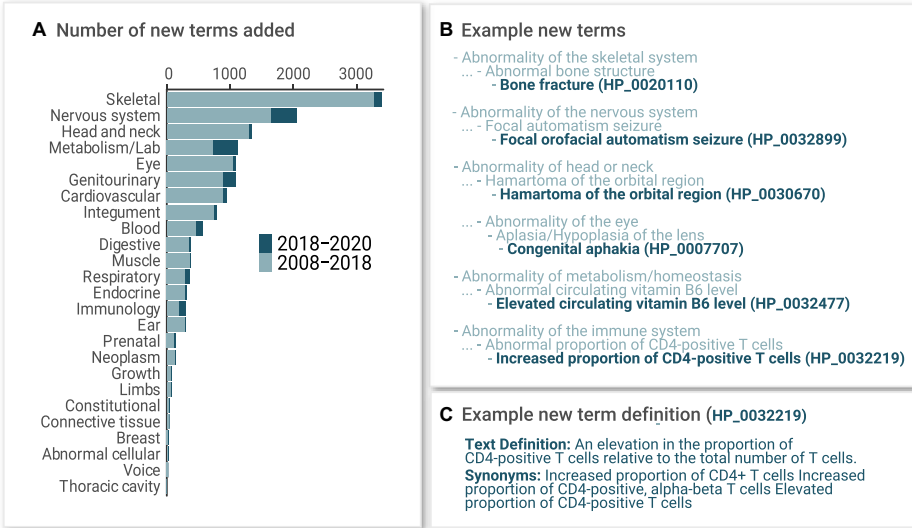
NEUROLOGIC
Central Nervous System
- Mental retardation
- Delayed development
- Delayed speech
- Spasticity
- Tremor
- Ataxia
- Parkinsonism
- Shuffling gait
- Spastic gait
- Hyperreflexia
- Increased tone

- 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)



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

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

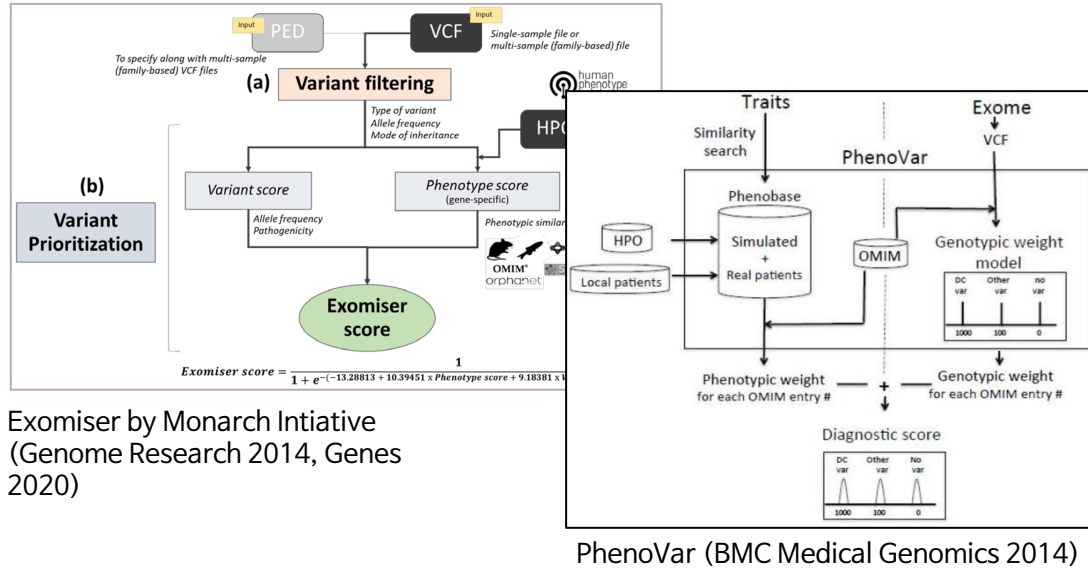
Digestive System [2 annotations]

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

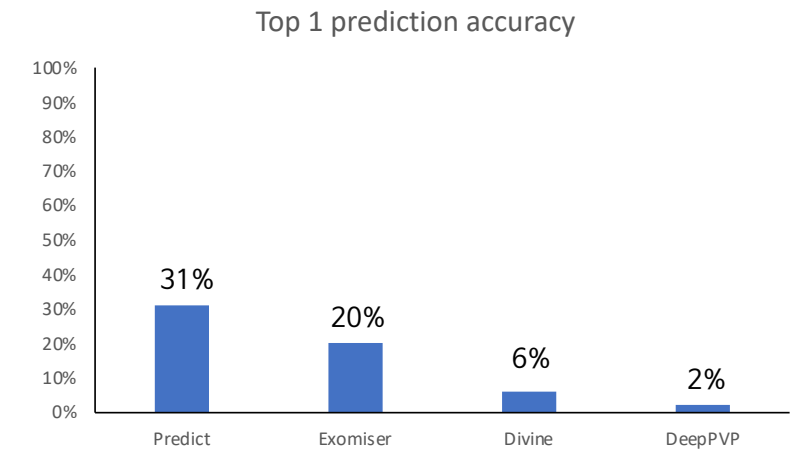
Skeletal system [1 annotation]

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

Advances in Utilizing Phenotype Information: Integrated Tools



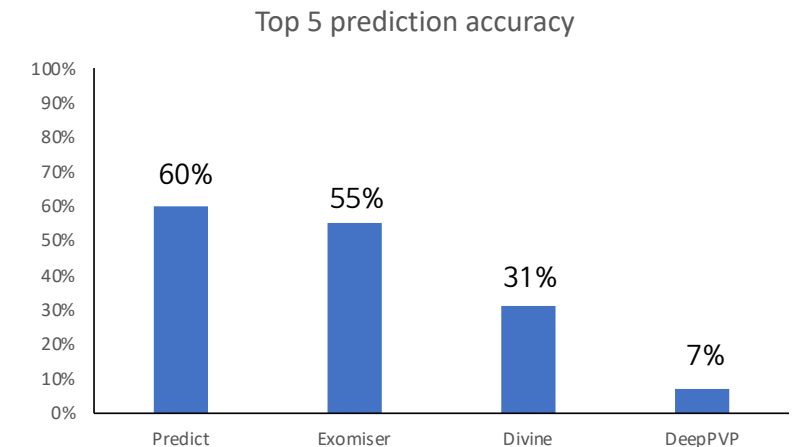
* Evaluation using the data of 108 patients with confirmed diagnosis (unpublished)



Divine (biorxiv 2018)

DeepPVP (BMC Bioinformatics 2019)

PREDICT (in preparation, beta)

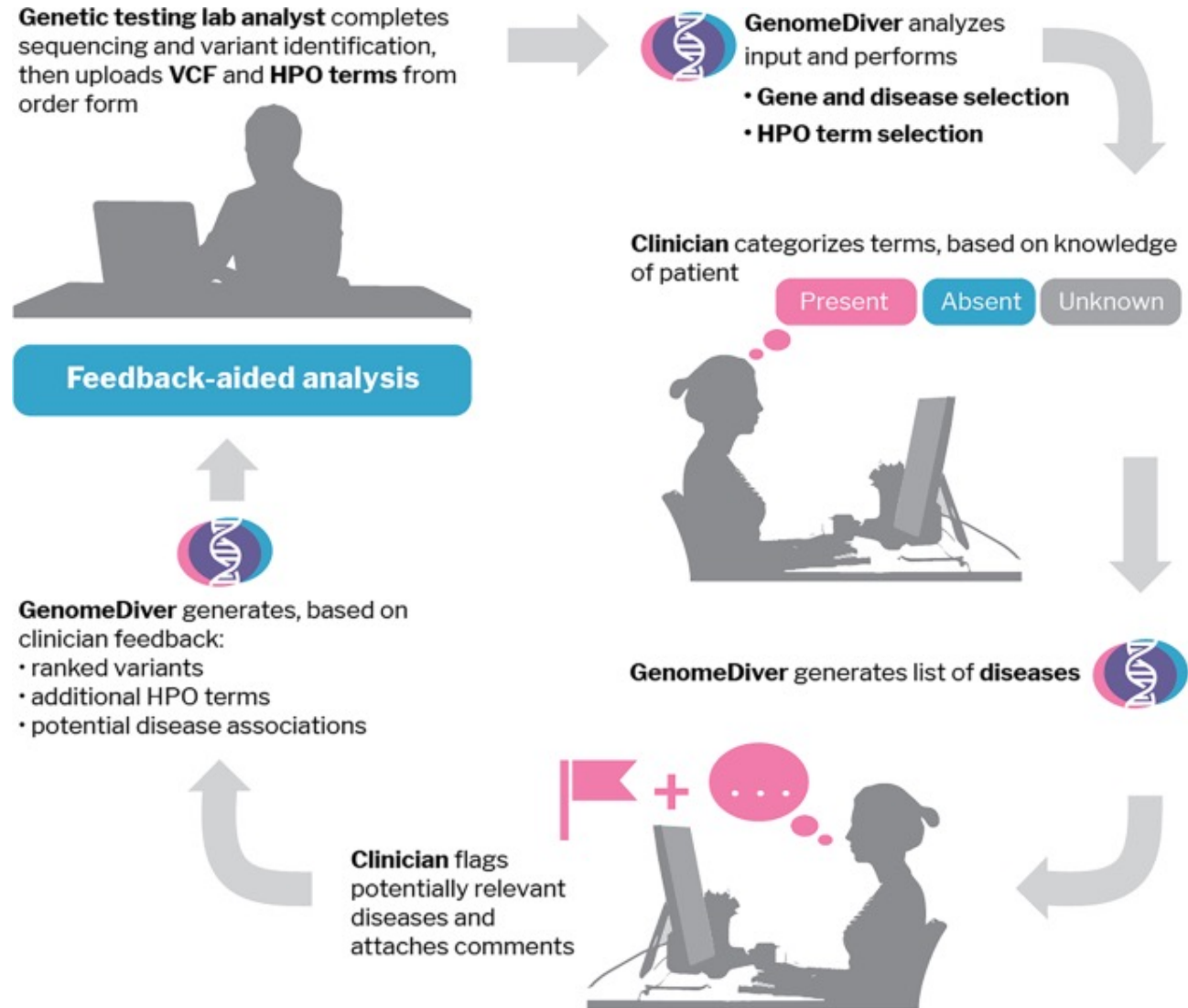


Application in Clinical Environment

Integration in clinical environment

• **Example) GenomeDiver** (Pearson et al., *Genetics in Medicine* 2021)

- 1) Patient's variants and initial phenotyping are put into the system.
- 2) The system generates candidate genes (and diseases) with relevant phenotypes.
- 3) Clinical refines patient's phenotypes.
- 4) The system re-analyzes using the refined phenotypes.
- 5) Feedback to data analyst.



Application in Clinical Environment: Regulations in Korea

Medical devices guideline (2017)

< 표 1. 의료영상을 이용한 빅데이터 및 인공지능 기술이 적용된 의료기기의 품목 예시 >

번호	품목명(등급)	정의
1	의료영상분석장치 소프트웨어(2)	의료영상을 획득하여 모의 치료, 모의 시술, 진단에 사용가능하도록 분석하는 장치에 사용하는 소프트웨어
2	방사선치료계획 소프트웨어(2)	획득된 의료용 영상을 이용하여 방사선 모의 치료 및 모의 시술에 사용되는 소프트웨어
3	의료영상검출보조 소프트웨어(2)	의료영상 내에서 정상과 다른 이상 부위를 검출 한 후 윤곽선, 색상 또는 지시선 등으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어
4	의료영상진단보조 소프트웨어(3)	의료영상을 사용하여 질병의 유무, 질병의 중증도 또는 질병의 상태 등에 대한 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어

< 표 2. 의료영상 이외의 의료정보를 이용한 빅데이터 및 인공지능 기술이 적용된 의료기기의 품목(안) >

번호	품목명(등급)	정의
1	생체신호검출보조 소프트웨어(2)	환자의 각종 생체정보(의료영상 제외)를 사용하여 정상과 다른 이상 신호를 검출한 후 알람을 제공하거나 색상 또는 지시선 등으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어
2	생체신호진단보조 소프트웨어(3)	환자의 각종 생체정보(의료영상 제외)를 사용하여 질병의 유무, 질병의 중증도 또는 질병의 상태 등을 진단 또는 예측하거나 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어
3	인체유래검체 검출보조 소프트웨어(2)	인체 유래 검체를 분석하여 정상과 다른 특이적인 결과를 제공하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어
4	인체유래검체 진단보조 소프트웨어(3)	인체 유래 검체를 분석하여 질병의 유무, 질병의 중증도 또는 질병의 상태 등을 진단 또는 예측하거나 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어

Law on IVD (2020. 5. 1.)

체외진단의료기기법

[시행 2020. 5. 1.] [법률 제16433호, 2019. 4. 30., 제정]

식품의약품안전처(의료기기정책과), 043-719-3755

제1장 총칙

○ 제1조(목적) 이 법은 체외진단의료기기의 제조·수입 등 취급과 관리 및 지원에 필요한 사항을 규정하여 체외진단의료기기의 안전성 확보 및 품질 향상을 도모하고 체외진단의료기기의 국제경쟁력을 강화함으로써 국민보건 향상 및 체외진단의료기기의 발전에 이바지함을 목적으로 한다.

○ 제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.

1. "체외진단의료기기"란 사람이나 동물로부터 유래하는 검체를 체외에서 검사하기 위하여 단독 또는 조합하여 사용되는 시약, 대조·보정 물질, 기구·기계·장치, 소프트웨어 등 「의료기기법, 제2조제1항에 따른 의료기기로서 다음 각 목의 어느 하나에 해당하는 제품을 말한다.
 - 가. 생리학적 또는 병리학적 상태를 진단할 목적으로 사용되는 제품
 - 나. 질병의 소인(素因)을 판단하거나 질병의 예후를 관찰하기 위한 목적으로 사용되는 제품
 - 다. 선천적인 장애에 대한 정보 제공을 목적으로 사용되는 제품
 - 라. 혈액, 조직 등을 다른 사람에게 수혈하거나 이식하고자 할 때 안전성 및 적합성 판단에 필요한 정보 제공을 목적으로 사용되는 제품
 - 마. 치료 반응 및 치료 결과를 예측하기 위한 목적으로 사용되는 제품
 - 바. 치료 방법을 결정하거나 치료 효과 또는 부작용을 모니터링하기 위한 목적으로 사용되는 제품
2. "검체"란 인체 또는 동물로부터 수집하거나 채취한 조직·세포·혈액·체액·소변·분변 등과 이들로부터 분리된 혈청, 혈장, 염색체, DNA(Deoxyribonucleic acid), RNA(Ribonucleic acid), 단백질 등을 말한다.
3. "임상적 성능시험"이란 체외진단의료기기의 성능을 증명하기 위하여 검체를 분석하여 임상적·생리적·병리학적 상태와 관련된 결과를 확인하는 시험을 말한다.

P01000 체외진단 소프트웨어 IVD software for diagnosis

P01010.01 질환진단검사소프트웨어 [2] IVD software for diagnosis to disease, except cancer, tumor 다수의 임상검사정보만 입력하거나 생체지표를 추가 입력하여 감염진단, 미생물 동정, 특정질환(암 제외)의 진단보조 등 진단정보를 제공하는 소프트웨어

Application in Clinical Environment: Regulations in Korea

Target disease selection

- IVD approval requires clinical performance test.
- Clinical performance test covering all rare genetic disorders – Very difficult

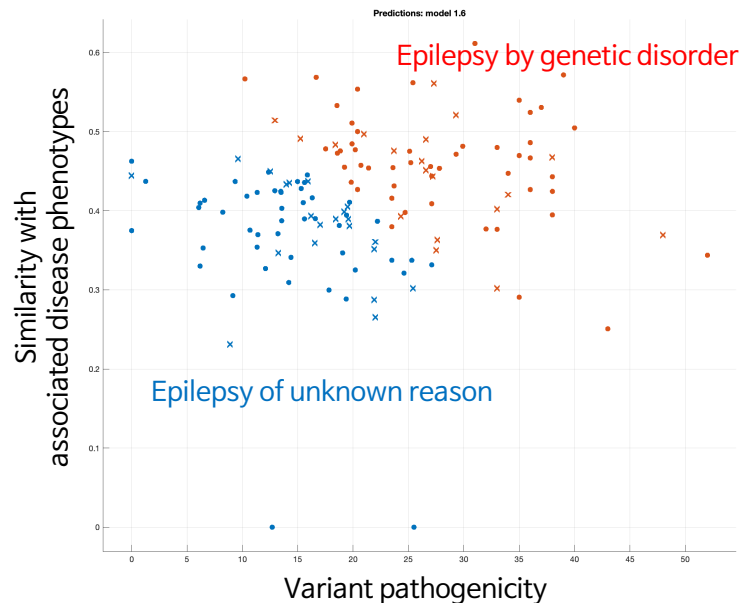
GMP Certification

- With a product, product manufacturing system, and quality management system

Clinical performance test

- Typical rare genetic disorder diagnosis software: 2nd grade IVD software
- Requires: Approval of plan for clinical performance test
- Requires: Clinical performance test at MFDS-designated test sites
- Prospective test can be very difficult for rare genetic disorders. – Retrospective test may be used.
- ... maybe 1st IVD software for rare genetic disorder?

Diagnosis of epilepsy by genetic disorder



Summary

- ◆ **Informatics nature in 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.

- ◆ **Ongoing changes in laws and regulations**
 - For better application of informatics technologies