Developing an Approach to Diagnose Neurodevelopmental Disorders based on Integrated Data Analysis

Sungwon Jung, Ph.D.
Gachon University College of Medicine
Gachon Institute of Genome Medicine and Science, Gachon University Gil Medical Center

Nov 1, 2018.
TBC/BIOINFO 2018

가천유전체의과학연구소
Gachon Institute of Genome Medicine and Science
嘉泉基因组医科研究院

가천대학교
Gachon University

Talk Outline

- **Neurodevelopmental disorders**
  - Characteristics and diagnosis
- **Efforts toward supporting tools**
- **Integrated analysis of variants, phenotypes, and MRI**
  - Overall concepts
  - Specifics
  - Preliminary performances

Most of this presentation is before publication

Neurodevelopmental Disorders

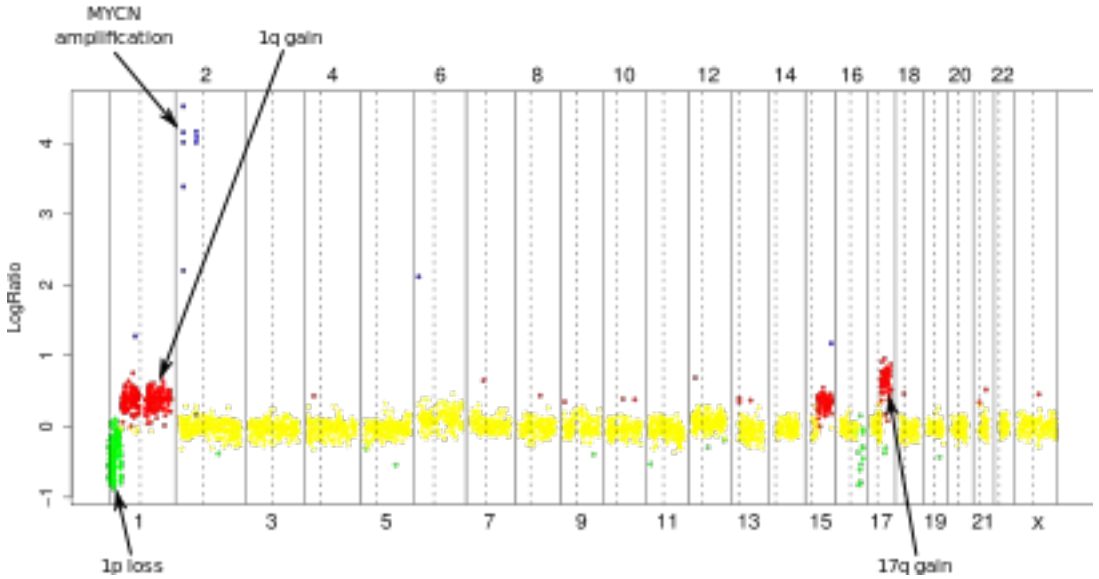
- **General definition: A group of disorders in which the development of the central nervous system is disturbed**

- **Causes**
  - Deprivation
  - **Genetic disorders** ← *Our focus*
  - Immune dysfunction
  - Infectious diseases
  - Metabolic disorders
  - Nutrition
  - Physical trauma

Our main target: Genetic disorders, with certain additional considerations for neurodevelopmental defects

Diagnosing Genetic Disorders

- **First line approach: Using microarrays for chromosome abnormalities and copy-number variants**
  - Diagnostic yield in about 20% of cases



- **The rest of cases often become undiagnosed patients.**
  - Searching (relatively) small pathogenic variants using whole exome/genome-sequencing (WES/WGS)

Diagnosing Genetic Disorders

- **A general approach**
  - Identifying disease-causing genetic variants
    - Identifying germline variants using Targeted-seq/WES/WGS
    - Prioritizing candidate variants
      - Previously reported with certain diseases?
      - Functional impact on proteins?
      - Matching allelic status with candidate diseases?
      - Rare in population?
  - Evaluating patient's phenotypes
    - Comparison with that of previously reported diseases
  - (If necessary) Identifying defects in brain development
    - Comparison with that of previously reported diseases
  - Sum it all for final diagnosis
- **Challenges**
  - RED: Usually requires **many knowledge sources** or needs **expertise of well-trained clinicians**

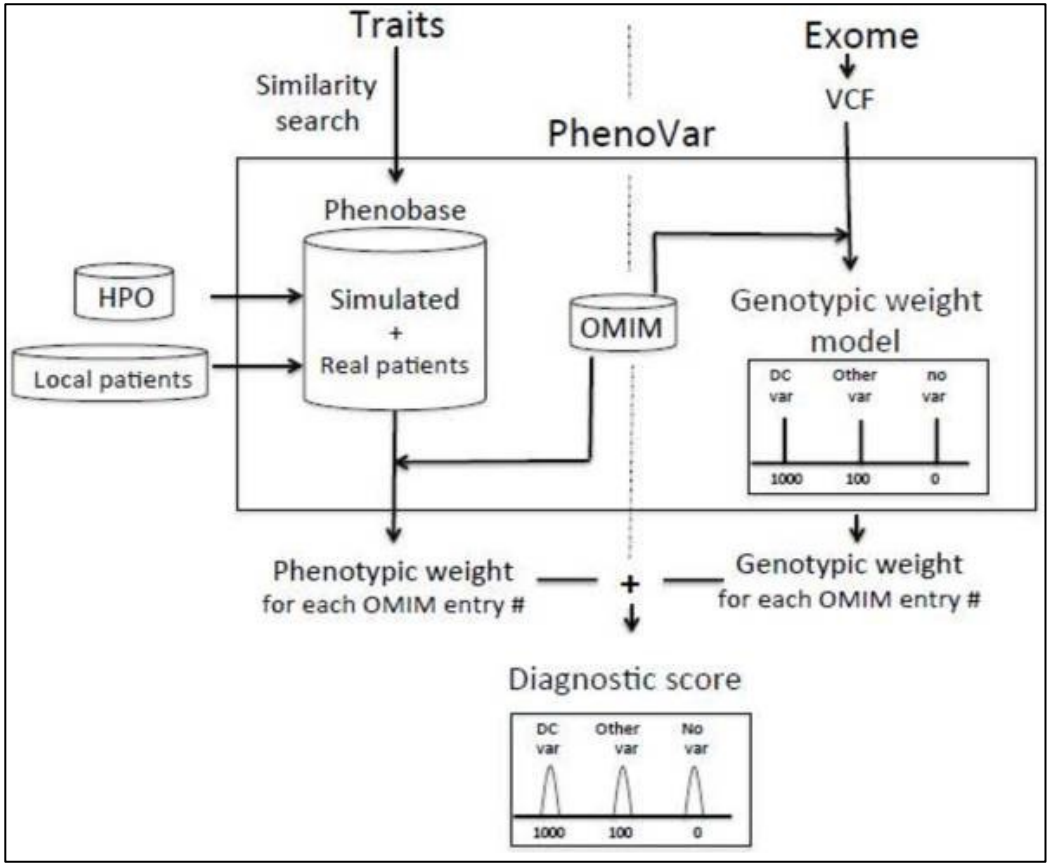
Examples of Supporting Tools

- **PhenoVar** [Trakadis et al., BMC Med Genomics 2014; Thuriot et al., Genetics in Medicine 2018]

Including simulated patients and real patients using HPO and OMIM databases

Twenty to twenty-five simulated patients for each syndrome list in OMIM entry

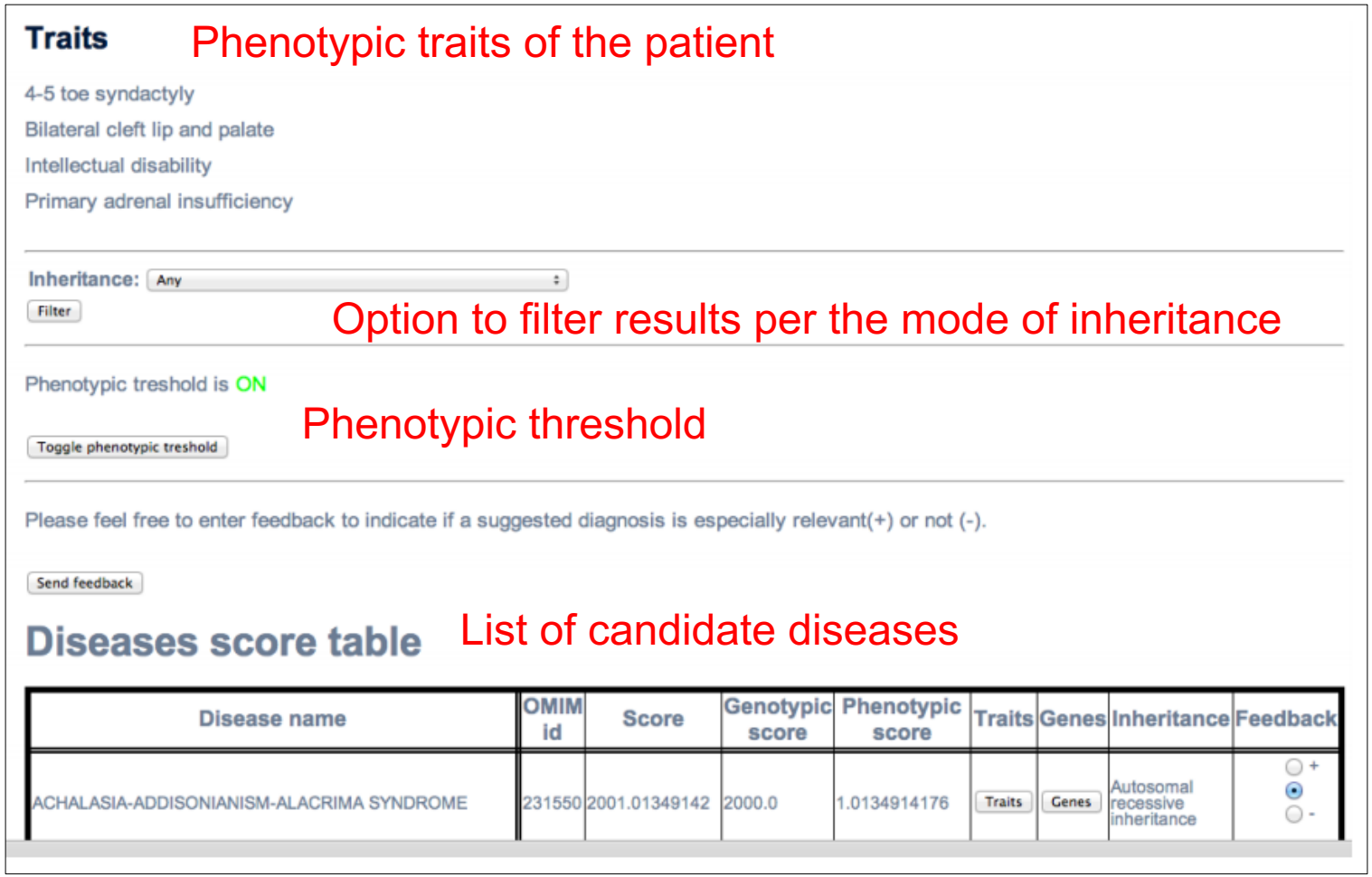
- 5 traits on average
- Modified VCF file including a pathogenic variant from the literature



Automatically prioritizes diagnoses based on phenotypes and genotypes

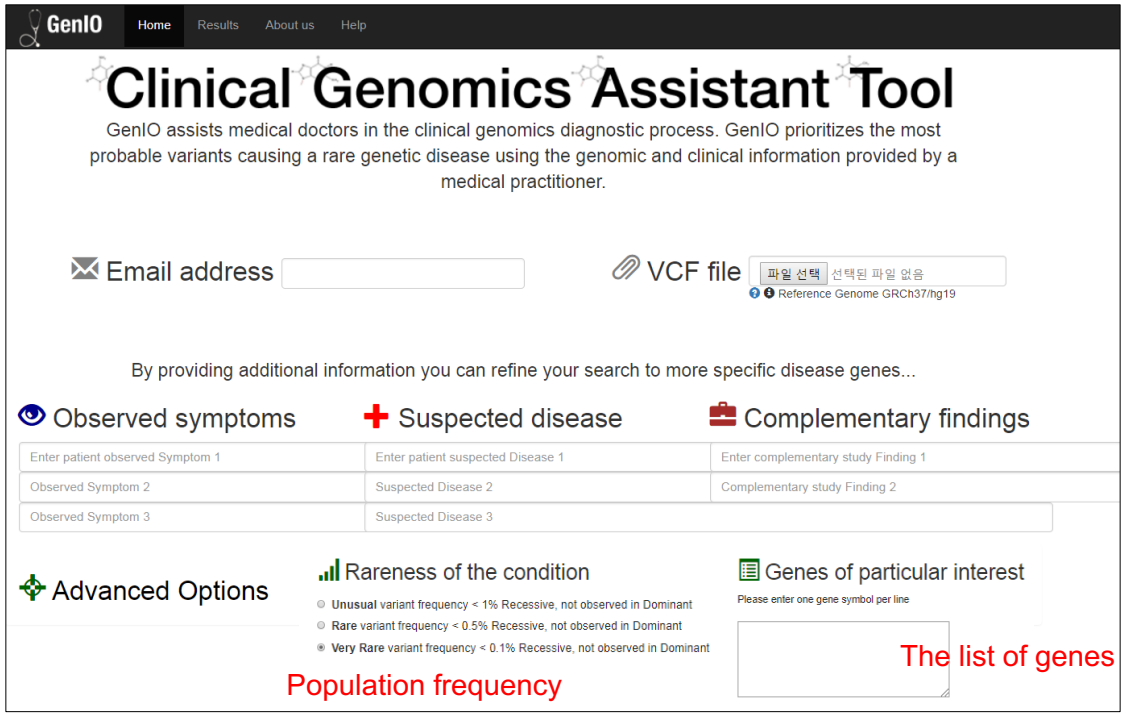
Examples of Supporting Tools

- **PhenoVar**



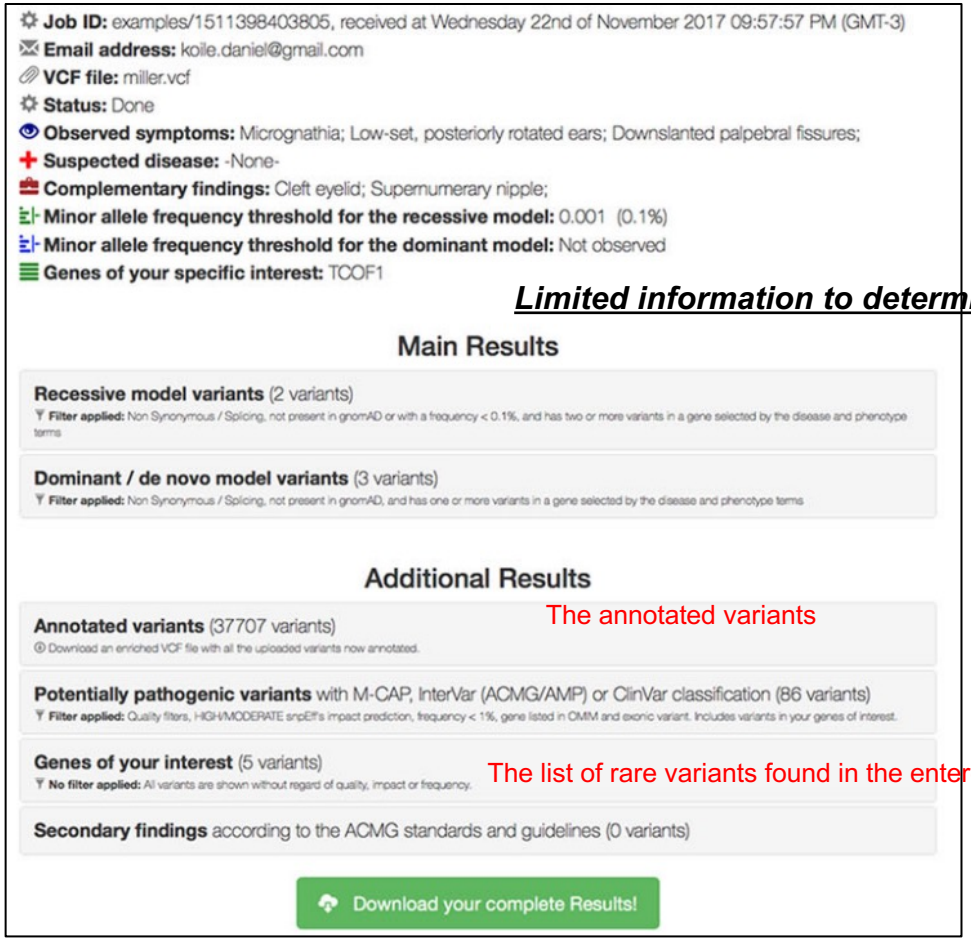
Examples of Supporting Tools

- **GenIO** [Koile et al., BMC Bioinformatics 2018]
  - Assisting clinicians to diagnose rare genetic diseases



Examples of Supporting Tools

- **GenIO**
  - Shows candidate variants



Limited information to determine (final) diagnosis

The list of rare variants according to mode of inheritance

The annotated variants

The list of potential pathogenic variants

The list of rare variants found in the entered list of genes of interest

Our Approach

- Most similar evidences
... Detailed data of evidences
(phenotype, disease causing variants, MRI)

Patient information
- Phenotype
- VCF
- MRI

Preprocessing

HPO

Annotation Filtering Prioritization

Image Preprocessing

Feature extraction

Similarity evaluation

Combined similarity

Phenotype similarity

Variant Matching

Brain feature Similarity

SNU cohort

Phenotype

Disease causing variant

MRI

ExAC
Esp6500
1000genome

OMIM
HGMD

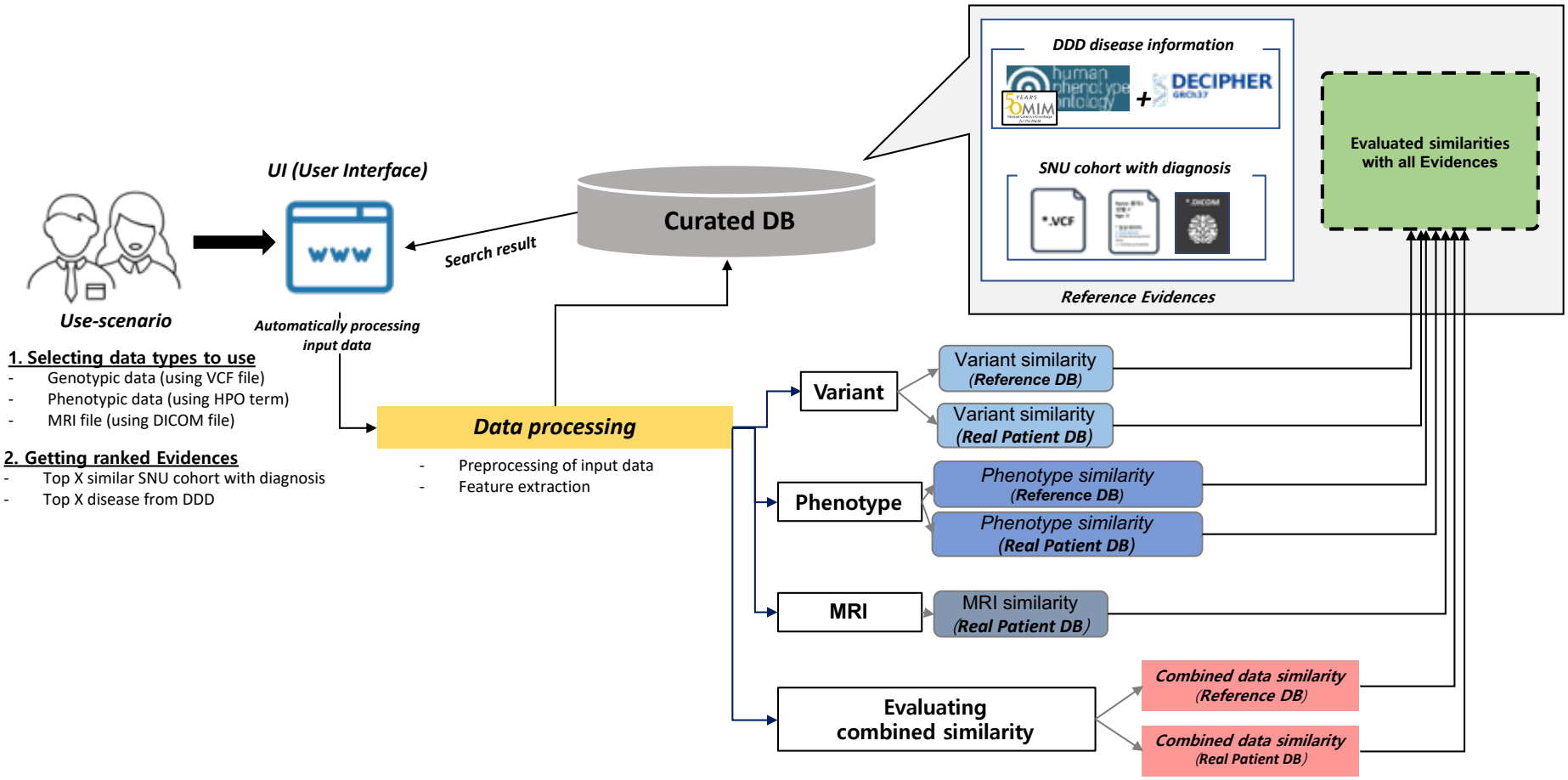
Annotation databases

DDD

Phenotype

Disease causing variant

System Structure



Processing Variants

Input VCF file

Annotation

ANNOVAR-based annotation

Filtering & prioritization

Filtering & prioritizing variants

- *Additional annotation with OMIM*
- *Frequency filtering*
- *Functional impact filtering*
- *Matching inheritance conditions & allelic status*

Pathogenic probability & rank evaluation

Probabilistic evaluation of pathogenic prediction annotations
Ranking variants

Similarity evaluation with Evidences

Rank-based similarity with disease genes in Evidences

Evaluating Pathogenic Probability of Variants

Annotated pathogenic predictions from ANNOVAR

V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
CLINSIG	CLNDBN	CLNACC	CLNDSDB	CLNDSDB	SIFT_scor	SIFT_pred	Polyphen2	Polyphen2	Polyphen2	Polyphen2	LRT_score	LRT_pred	Mutation	Mutation	Mutation	Mutation	FATHMM	FATHMM	R
Pathogenic	Immunodefi	RCV0001621	MedGen:OM	CN221808:61	0	D	.	.	.	.	0.352	N	1	D	.	.	.	.	.
Pathogenic	Immunodefi	RCV0001489	MedGen:OM	CN221808:61	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
Pathogenic	Immunodefi	RCV0001489	MedGen:OM	CN221808:61	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	0.34	T					0.204	N	1	A					

Q: How likely is a variant pathogenic given these predictions?

=> *P(variant = pathogenic | predictor A = a)*

By Bayes' theorem,

P(variant = pathogenic | predictor A = a)

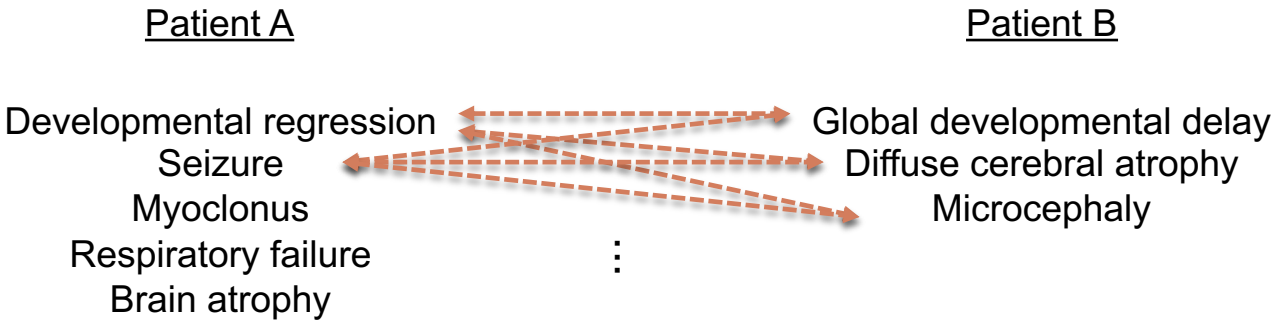
= *P(predictor A = a | variant = pathogenic)* x *P(variant = pathongeic)* / *P(predictor A = a)*

Based on the pathogenicity prediction of known pathogenic variants

Based on the statistics from the SNU cohort variants

Then, the probabilities of multiple predictions are aggregated.
Variants are ranked based on the aggregated pathogenic probability.

Evaluating Phenotype Similarity



Ontology-based term-to-term similarity:

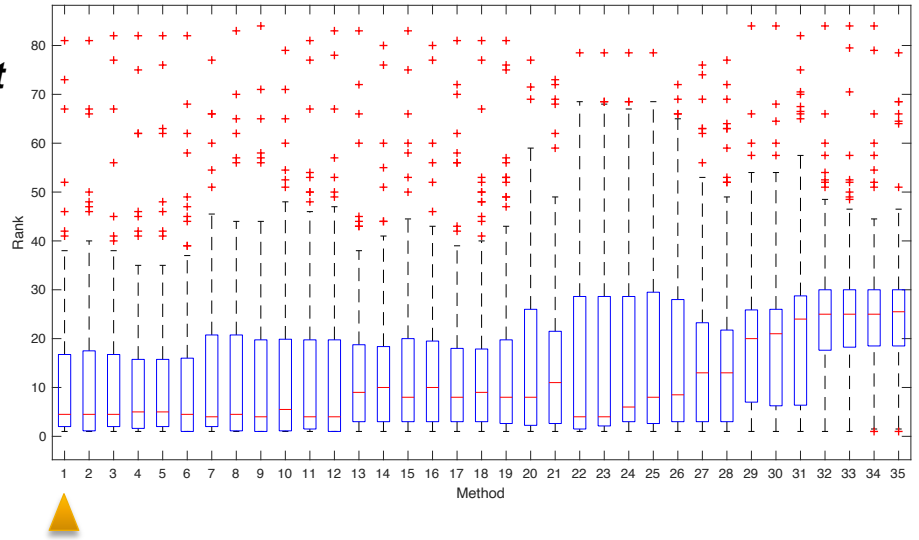
Information coefficient, Jiang-Conrath, Graph IC, Relevance, Wang, Lin, Resnik, ...

Aggregating multiple term-to-term similarities:

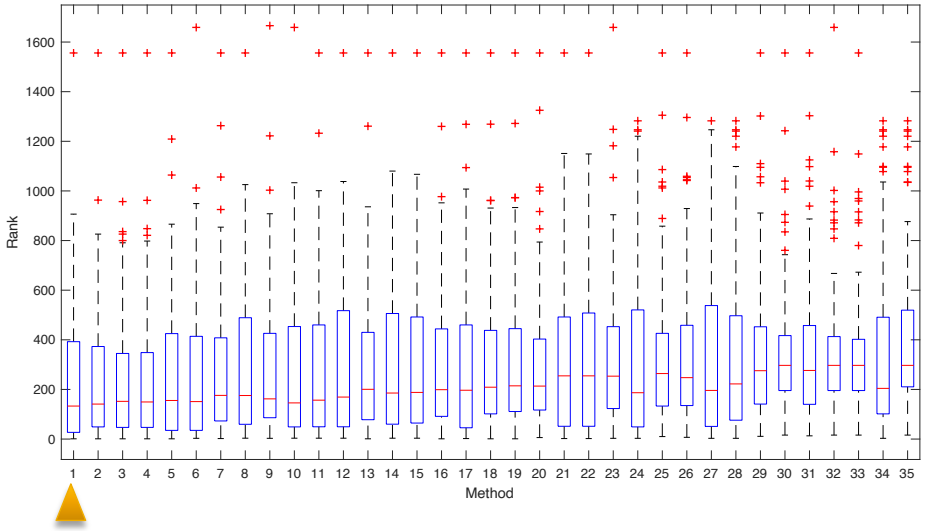
Max, Mean, FunSimAvg, FunSimMax, BMA, ...

Evaluating Phenotype Similarity

Similarity rank of
SNU cohort case ***to SNU cohort***
(LOOCV-like)



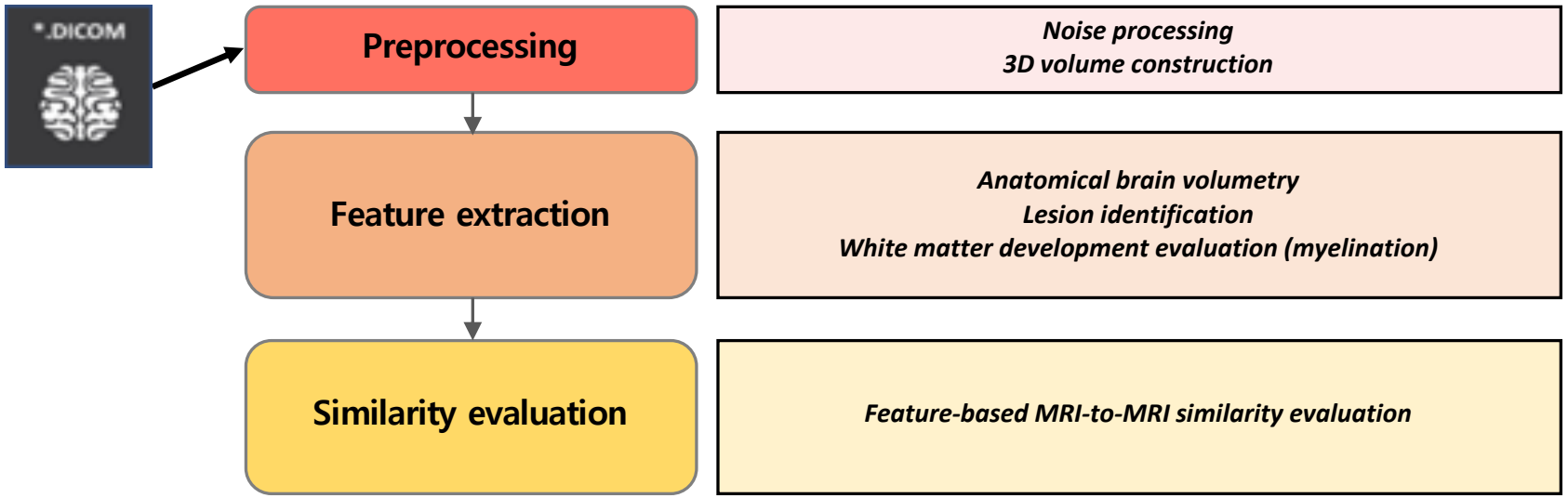
Similarity rank of
SNU cohort case ***to DDD***



**** Different term-to-term similarity method.***
Same aggregation method

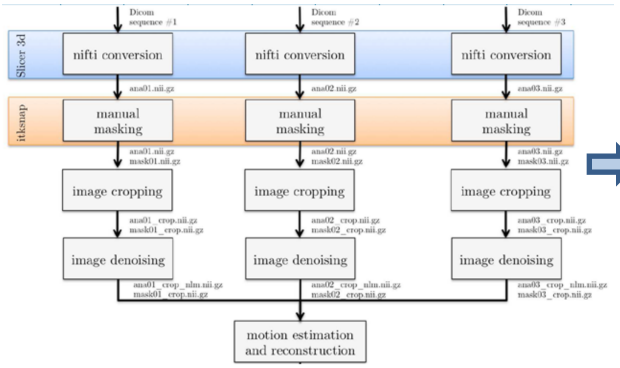
MRI Data Processing

Input MRI (DICOM) file

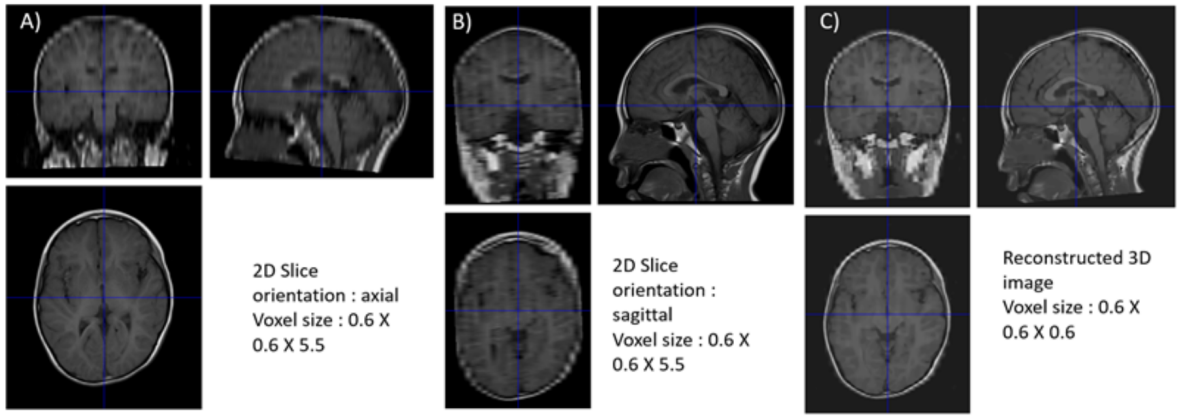


MRI Data Processing

Preprocessing: 3D volume construction

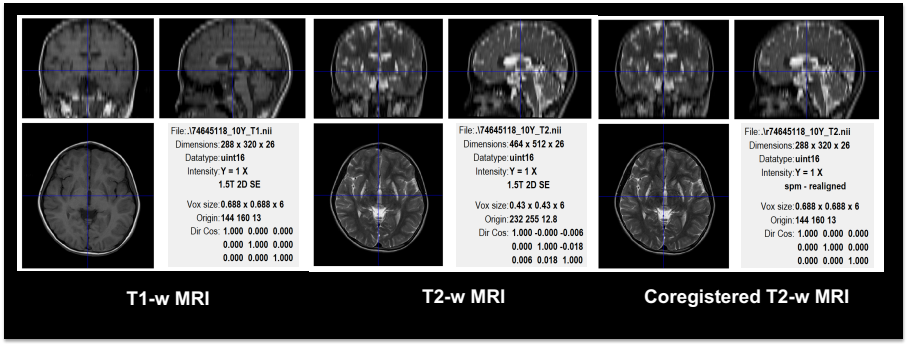


2D Superresolution pipeline



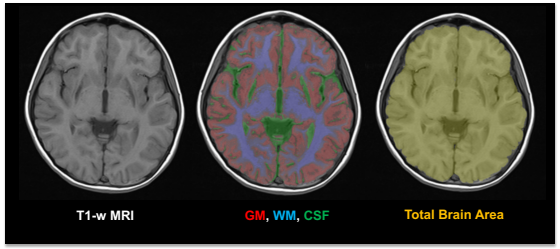
Original 2D image and reconstructed 3D brain image

Preprocessing: Overall image alignment

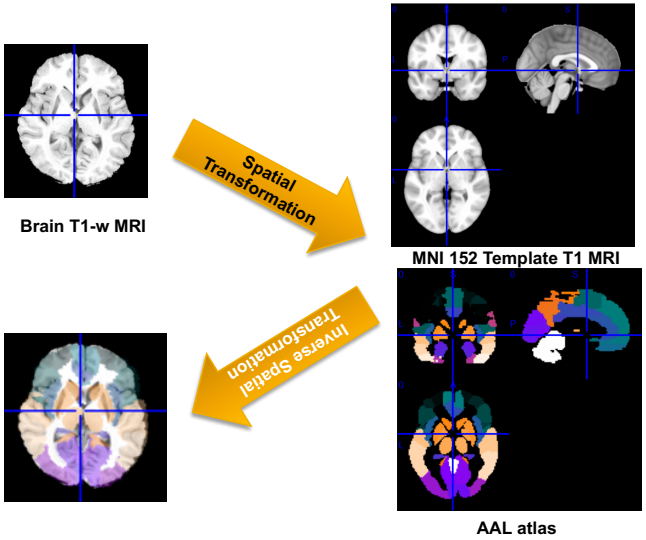


MRI Data Processing

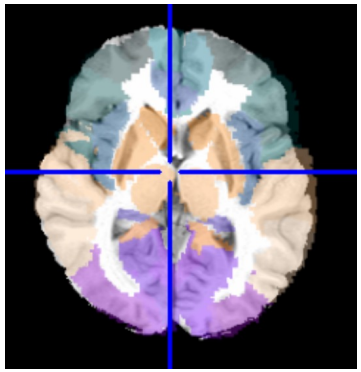
◆ Anatomical brain volumetry



Brain extraction



Anatomical region identification

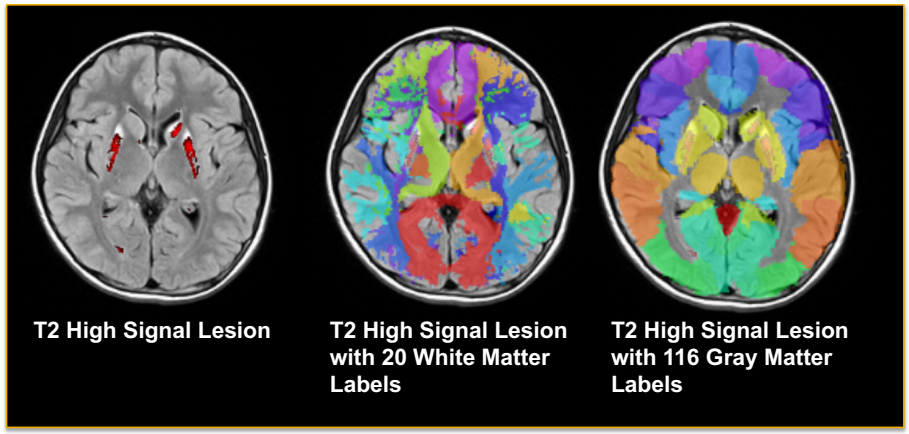


74645118_10Y	Volume in mm3
Total Brain Volume	1363827.985
Total Gray matter Volume	808412.4138
Total White matter Volume	448472.3592
Total CSF Volume	311366.1129
Volume of Frontal Lobe	240720.0677
Volume of Temporal Lobe	186647.2427
Volume of Parietal Lobe	134996.3086
Volume of Occipital Lobe	99955.46179
Volume of Basal Ganglia	18362.6969
Volume of Thalamus	10266.09464
Volume of Cerebellum	114038.7286
Volume of left precentral gyrus	17565.7984
Volume of right precentral gyrus	15104.20443
Volume of left superior frontal gyrus (dorsolateral)	18629.27505
Volume of right superior frontal gyrus (dorsolateral)	22267.78318
Volume of left superior frontal gyrus, orbital	4631.086339
Volume of right superior frontal gyrus, orbital	5039.461374

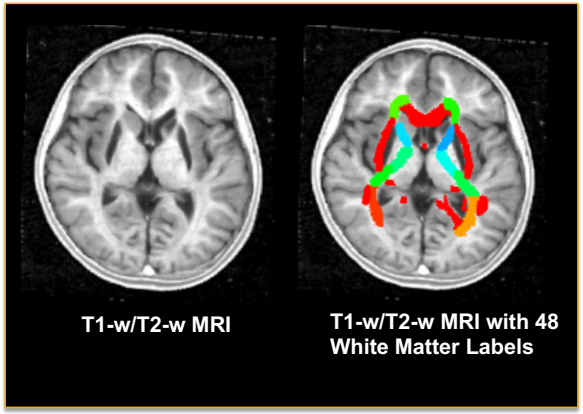
Anatomical volume evaluation

MRI Data Processing

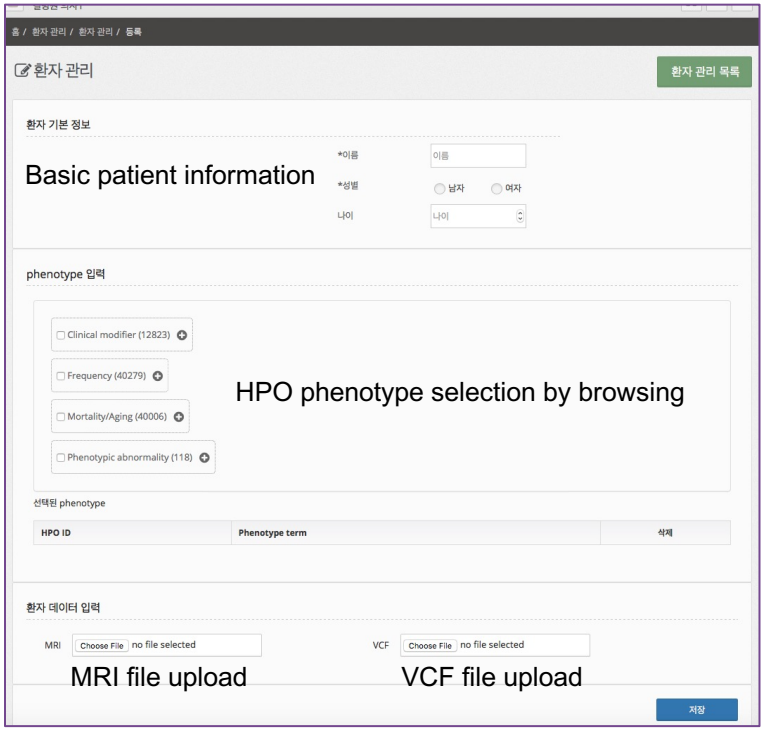
- **Lesion volumetry**



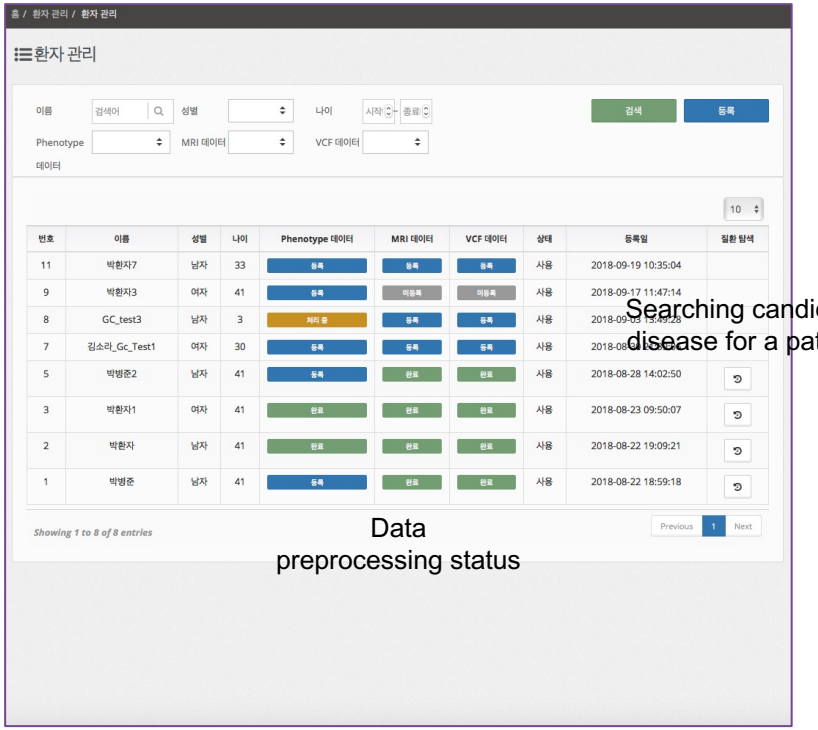
- **Myelination evaluation (related to white matter development)**



Software Interface under Development

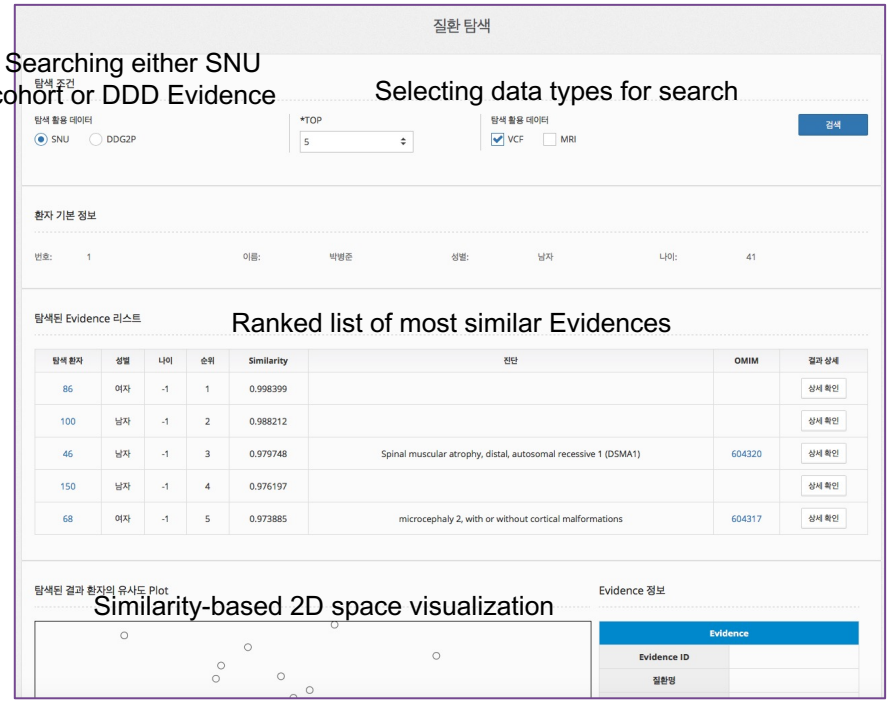


Input of new patient data

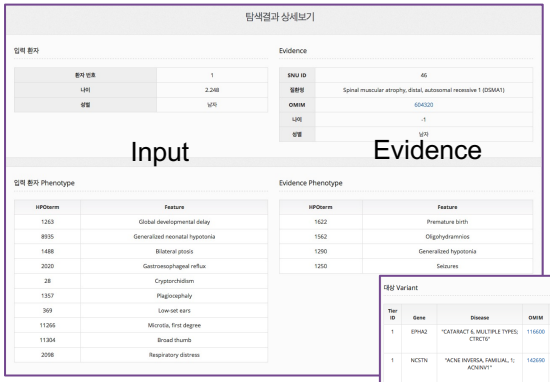


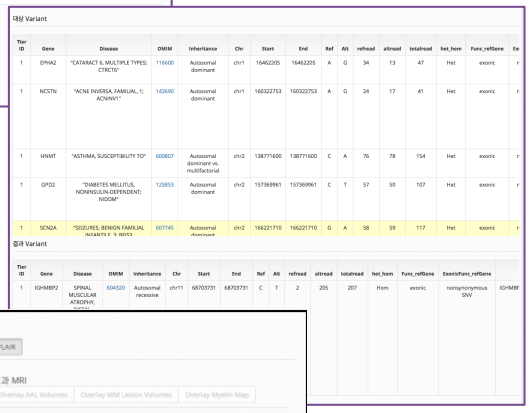
All the uploaded patients

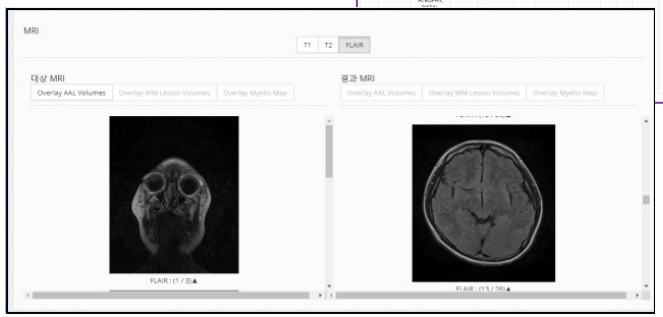
Software Interface under Development



Search result







Detailed side-by-side comparison with selected Evidence

Preliminary Performance Evaluation

- **Accuracy of comparing SNU cohort case to SNU cohort (like LOOCV)**
  - TOP 1 = true disease (without MRI, N = 31): 80.6%
  - TOP 1 = true disease (with MRI, N = 12): 83.3%
- **Accuracy of comparing SNU cohort case to DDD Evidence**
  - True disease within TOP 3 (N = 45): 77.8%
  - True disease within TOP 5 (N = 45): 86.7%
- **Divine (bioRxiv 2018, N = 26)**
  - Average rank of true disease: 5

  (2.7 by our system, to DDD Evidence, N = 45)
- **PhenoVar, Exomiser (Comparison in Thuriot et al., Genetics in Medicine 2018, N = 18)**
  - True disease within TOP 10
    - Exomiser: 56%
    - PhenoVar: 89%
    - *Our system (to DDD Evidence, N = 45)* *95.6%*