

# 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

# 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

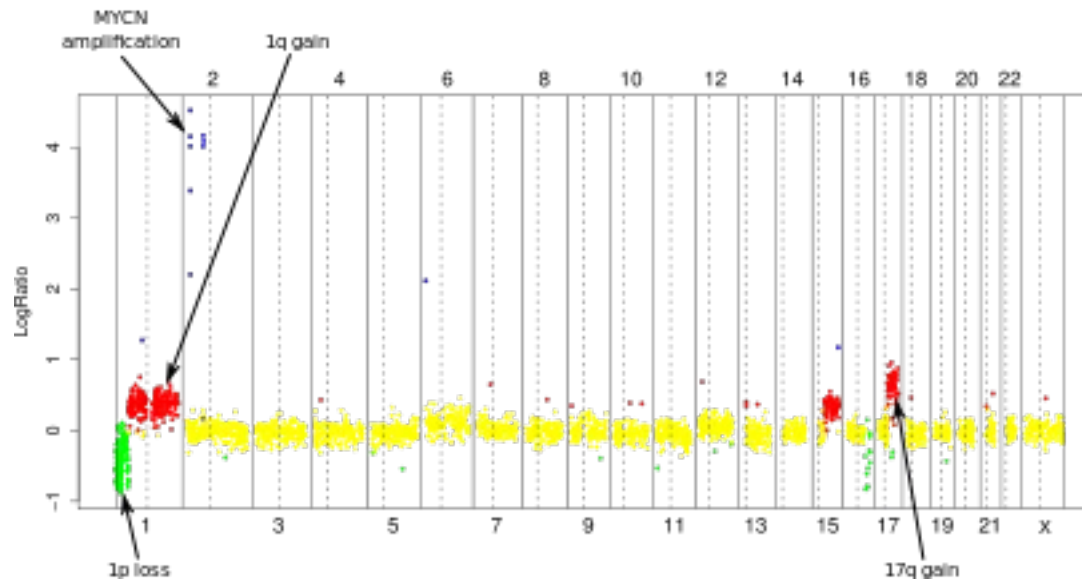
# Neurodevelopmental Disorders

- ◆ **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**

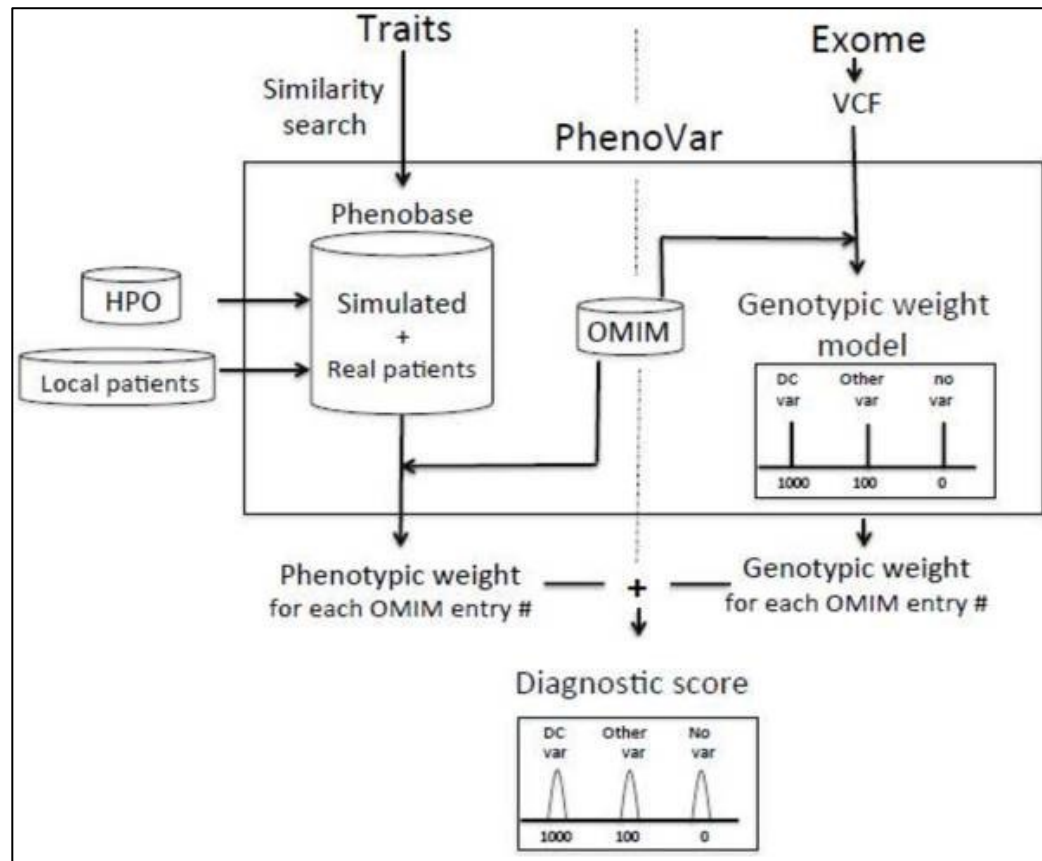
# Examples of Supporting Tools

◆ **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

**Traits**      Phenotypic traits of the patient

4-5 toe syndactyly  
Bilateral cleft lip and palate  
Intellectual disability  
Primary adrenal insufficiency

---

Inheritance:

     Option to filter results per the mode of inheritance

---

Phenotypic threshold is **ON**

     Phenotypic threshold

---

Please feel free to enter feedback to indicate if a suggested diagnosis is especially relevant(+) or not (-).

**Diseases score table**      List of candidate diseases

Disease name	OMIM id	Score	Genotypic score	Phenotypic score	Traits	Genes	Inheritance	Feedback
ACHALASIA-ADDISONIANISM-ALACRIMA SYNDROME	231550	2001.01349142	2000.0	1.0134914176	<input type="button" value="Traits"/>	<input type="button" value="Genes"/>	Autosomal recessive inheritance	<input type="radio"/> + <input checked="" type="radio"/> <input type="radio"/> -

# Examples of Supporting Tools

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

**GenIO** Home Results About us Help

## Clinical Genomics Assistant Tool

GenIO assists medical doctors in the clinical genomics diagnostic process. GenIO prioritizes the most probable variants causing a rare genetic disease using the genomic and clinical information provided by a medical practitioner.

✉ Email address

📎 VCF file  선택된 파일 없음  
 Reference Genome GRCh37/hg19

By providing additional information you can refine your search to more specific disease genes...

👁 Observed symptoms	⊕ Suspected disease	📄 Complementary findings
<input type="text" value="Enter patient observed Symptom 1"/>	<input type="text" value="Enter patient suspected Disease 1"/>	<input type="text" value="Enter complementary study Finding 1"/>
<input type="text" value="Observed Symptom 2"/>	<input type="text" value="Suspected Disease 2"/>	<input type="text" value="Complementary study Finding 2"/>
<input type="text" value="Observed Symptom 3"/>	<input type="text" value="Suspected Disease 3"/>	<input type="text"/>

🌿 **Advanced Options**

📊 **Rareness of the condition**

- Unusual variant frequency < 1% Recessive, not observed in Dominant
- Rare variant frequency < 0.5% Recessive, not observed in Dominant
- Very Rare variant frequency < 0.1% Recessive, not observed in Dominant

📄 **Genes of particular interest**  
Please enter one gene symbol per line

**The list of genes**

**Population frequency**

# Examples of Supporting Tools

- ◆ GenIO
  - Shows candidate variants

⚙ **Job ID:** examples/1511398403805, received at Wednesday 22nd of November 2017 09:57:57 PM (GMT-3)

✉ **Email address:** koile.daniel@gmail.com

📎 **VCF file:** miller.vcf

⚙ **Status:** Done

👁 **Observed symptoms:** Micrognathia; Low-set, posteriorly rotated ears; Downslanted palpebral fissures;

➕ **Suspected disease:** -None-

📖 **Complementary findings:** Cleft eyelid; Supernumerary nipple;

📊 **Minor allele frequency threshold for the recessive model:** 0.001 (0.1%)

📊 **Minor allele frequency threshold for the dominant model:** Not observed

📖 **Genes of your specific interest:** TCOF1

**Limited information to determine (final) diagnosis**

### Main Results

**Recessive model variants (2 variants)**

⚙ Filter applied: Non Synonymous / Splicing, not present in gnomAD or with a frequency < 0.1%, and has two or more variants in a gene selected by the disease and phenotype terms

**Dominant / de novo model variants (3 variants)**

⚙ Filter applied: Non Synonymous / Splicing, not present in gnomAD, and has one or more variants in a gene selected by the disease and phenotype terms

### Additional Results

**Annotated variants (37707 variants)**

📎 Download an enriched VCF file with all the uploaded variants now annotated.

**Potentially pathogenic variants with M-CAP, InterVar (ACMG/AMP) or ClinVar classification (86 variants)**

⚙ Filter applied: Quality filters, HIGH/MODERATE snPEff's impact prediction, frequency < 1%, gene listed in OMM and exonic variant. Includes variants in your genes of interest.

**Genes of your interest (5 variants)**

⚙ No filter applied: All variants are shown without regard of quality, impact or frequency.

**Secondary findings** according to the ACMG standards and guidelines (0 variants)

**Download your complete Results!**

The list of rare variants according to mode of inheritance

**Limited information to determine (final) diagnosis**

The annotated variants

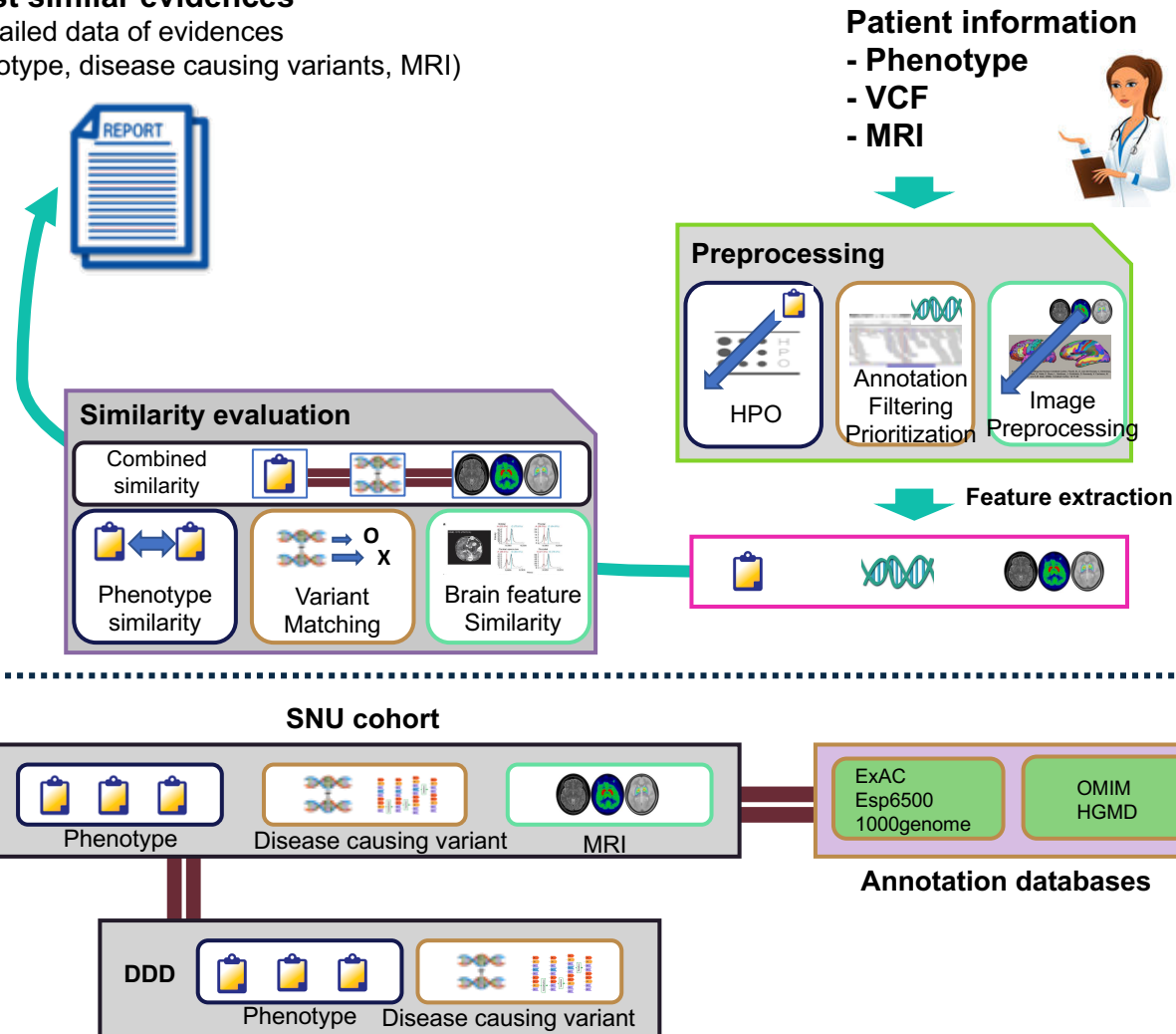
The list of potential pathogenic variants

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

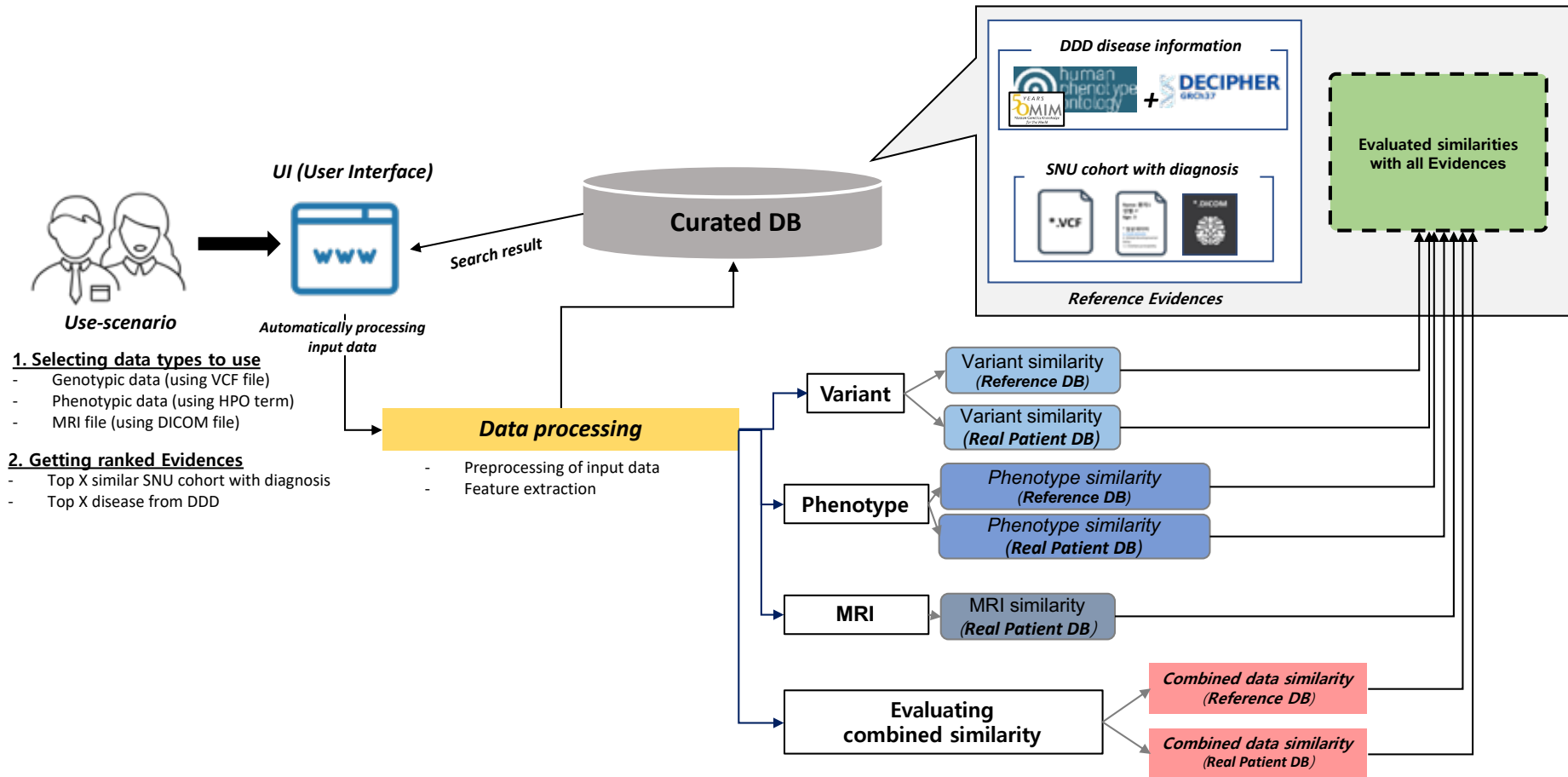
# Our Approach

## - Most similar evidences

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

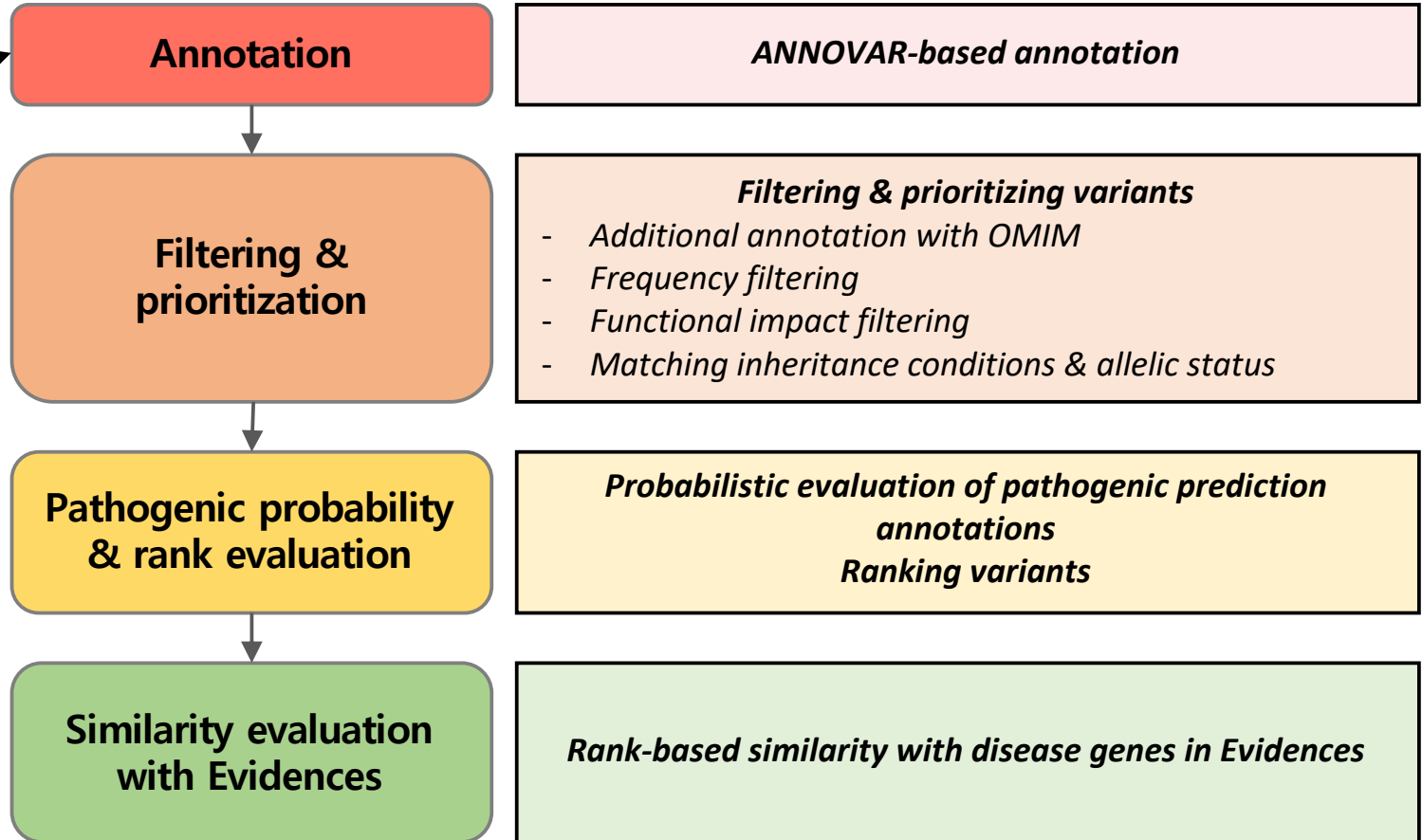


# System Structure



# Processing Variants

Input VCF file





# 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_score	SIFT_pred	Polyphen2	Polyphen2	Polyphen2	Polyphen2	LRT_score	LRT_pred	Mutation	Mutation	Mutation	Mutation	FATHMM	FATHMM
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
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	.	.	.	.

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

=>  $P(\text{variant} = \text{pathogenic} \mid \text{predictor } A = a)$

**By Bayes' theorem,**

$P(\text{variant} = \text{pathogenic} \mid \text{predictor } A = a)$

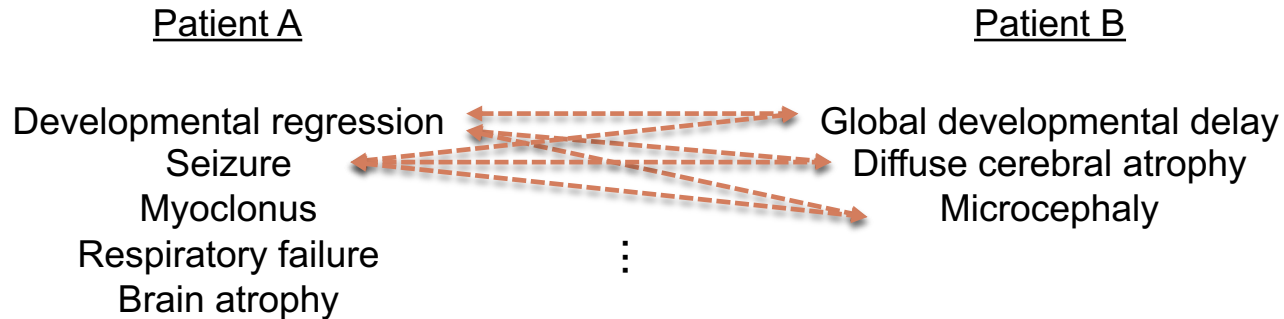
$= \frac{P(\text{predictor } A = a \mid \text{variant} = \text{pathogenic}) \times P(\text{variant} = \text{pathogenic})}{P(\text{predictor } A = a)}$

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:

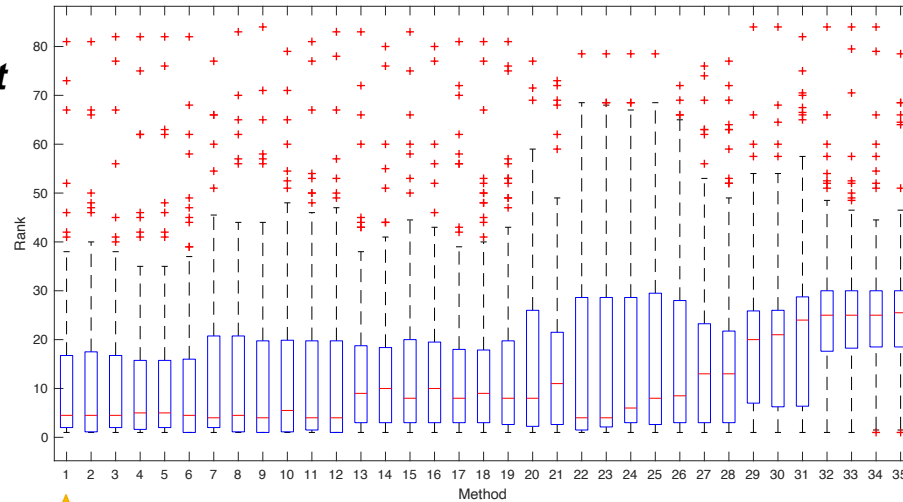
Information coefficient, Jiang-Conrath, Graph IC, Relevance, Wang, Lin, Resnik, ...

## Aggregating multiple term-to-term similarities:

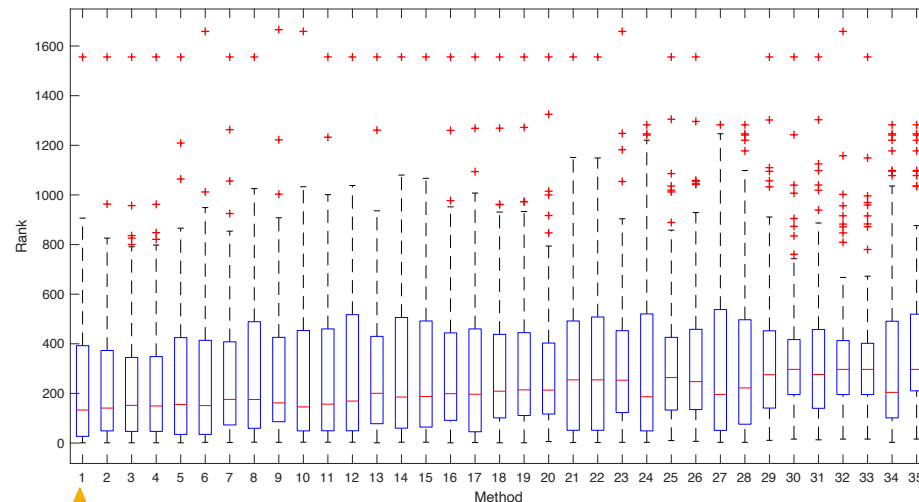
Max, Mean, FunSimAvg, FunSimMax, BMA, ...

# Evaluating Phenotype Similarity

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



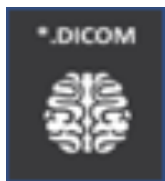
Similarity rank of  
SNU cohort case *to DDD*



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

# MRI Data Processing

Input MRI  
(DICOM) file



**Preprocessing**

*Noise processing*  
*3D volume construction*

**Feature extraction**

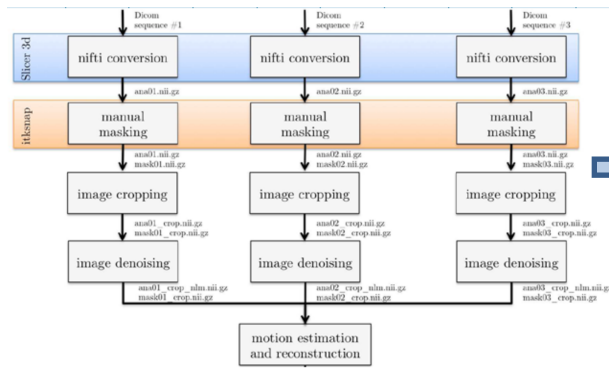
*Anatomical brain volumetry*  
*Lesion identification*  
*White matter development evaluation (myelination)*

**Similarity evaluation**

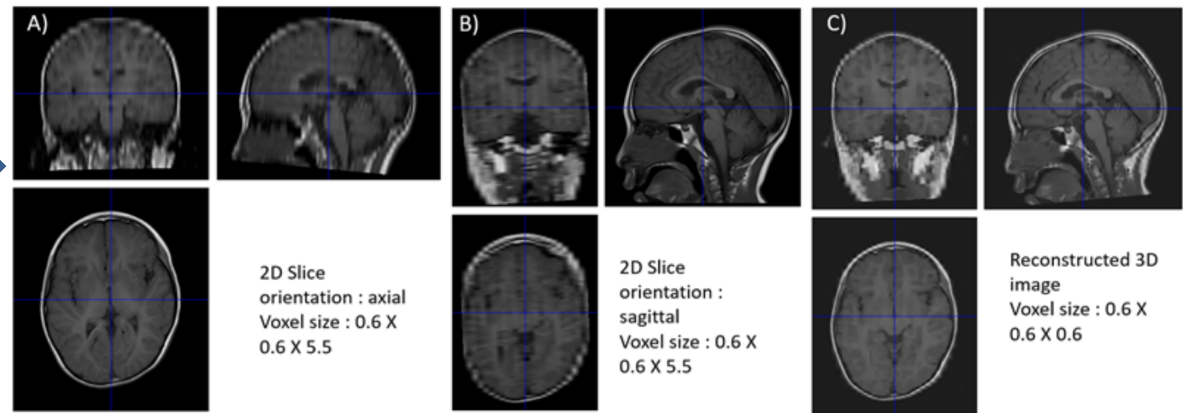
*Feature-based MRI-to-MRI similarity evaluation*

# MRI Data Processing

## ◆ Preprocessing: 3D volume construction

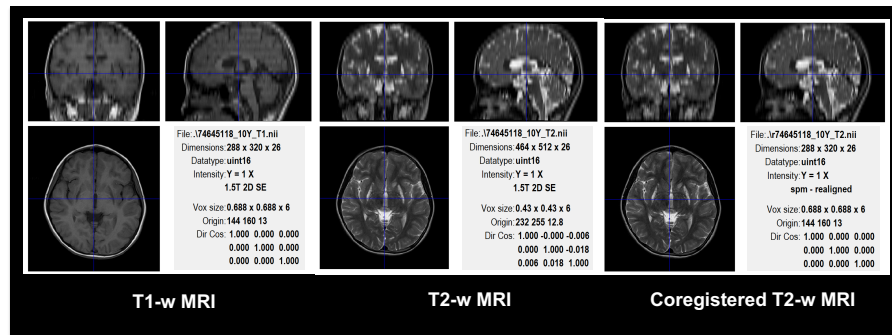


2D Superresolution pipeline



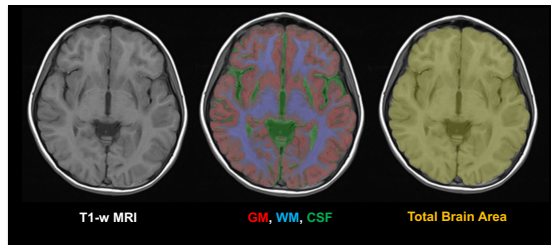
Original 2D image and reconstructed 3D brain image

## ◆ Preprocessing: Overall image alignment

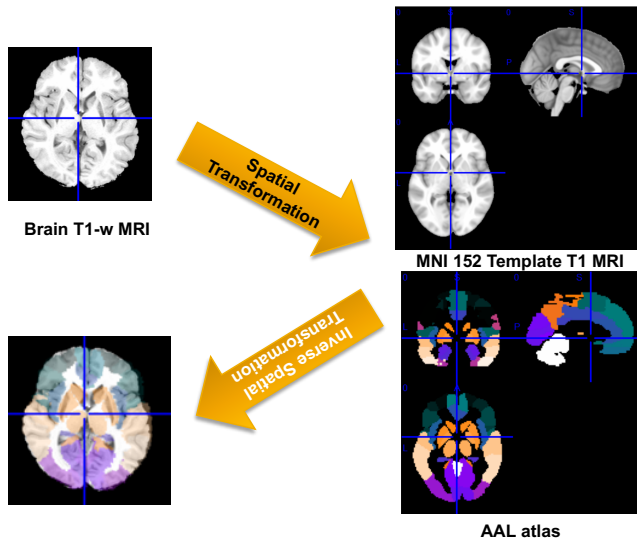
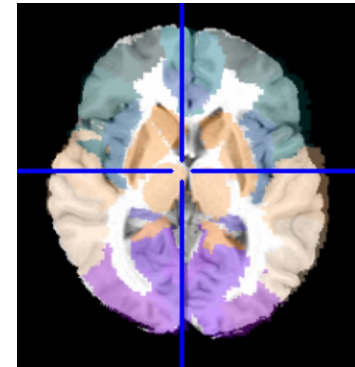


# MRI Data Processing

## ◆ Anatomical brain volumetry



Brain extraction



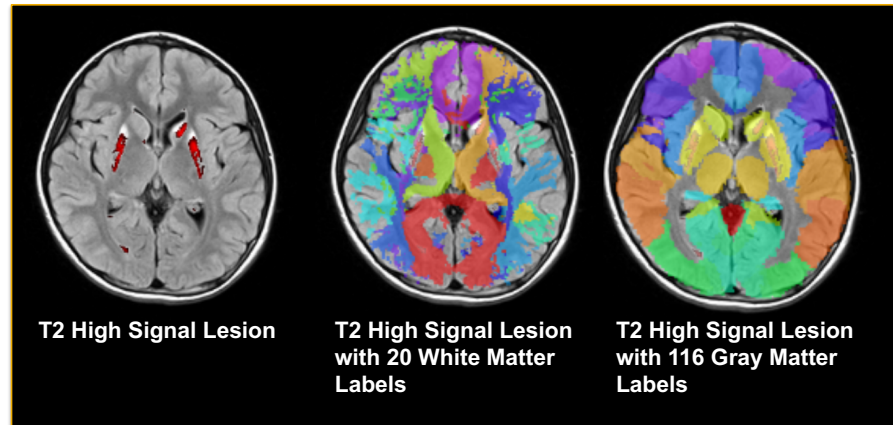
Anatomical region identification

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

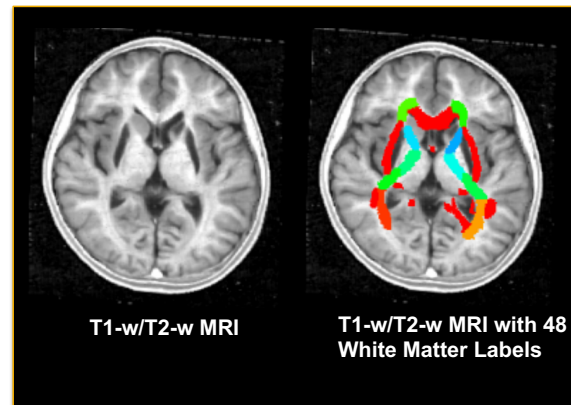
Anatomical volume evaluation

# MRI Data Processing

## ◆ Lesion volumetry



## ◆ Myelination evaluation (related to white matter development)



# Software Interface under Development

환자 관리

환자 기본 정보

Basic patient information

\*이름: 이름

\*성별:  남자  여자

나이: 나이

phenotype 입력

HPO phenotype selection by browsing

선택된 phenotype

HPO ID: Phenotype term: 삭제

환자 데이터 입력

MRI: Choose File no file selected

VCF: Choose File no file selected

MRI file upload

VCF file upload

저장

Input of new patient data

환자 관리

이름 검색어 Q 성별 나이 시작 종료

Phenotype MRI 데이터 VCF 데이터

데이터

번호	이름	성별	나이	Phenotype 데이터	MRI 데이터	VCF 데이터	상태	등록일	질환 탐색
11	박환자7	남자	33	등록	등록	등록	사용	2018-09-19 10:35:04	
9	박환자3	여자	41	등록	미등록	미등록	사용	2018-09-17 11:47:14	
8	GC_test3	남자	3	처리 중	등록	등록	사용	2018-09-04 13:49:28	
7	김소라_Gc_Test1	여자	30	등록	등록	등록	사용	2018-08-22 19:09:21	
5	박병준2	남자	41	등록	완료	완료	사용	2018-08-28 14:02:50	
3	박환자1	여자	41	완료	완료	완료	사용	2018-08-23 09:50:07	
2	박환자	남자	41	완료	완료	완료	사용	2018-08-22 19:09:21	
1	박병준	남자	41	등록	완료	완료	사용	2018-08-22 18:59:18	

Showing 1 to 8 of 8 entries

Data preprocessing status

Previous 1 Next

Searching candidate disease for a patient

All the uploaded patients



# Software Interface under Development

**질환 탐색**

Searching either SNU cohort or DDD Evidence  
 Selecting data types for search

탐색 기준: SNU (선택됨) / DDG2P  
 \*TOP: 5  
 탐색 활용 데이터:  VCF  MRI

환자 기본 정보  
 번호: 1 이름: 박병준 성별: 남자 나이: 41

탐색된 Evidence 리스트  
**Ranked list of most similar Evidences**

탐색 환자	성별	나이	순위	Similarity	증상	OMIM	결과 상세
86	여자	-1	1	0.998399			상세 확인
100	남자	-1	2	0.988212			상세 확인
46	남자	-1	3	0.979748	Spinal muscular atrophy, distal, autosomal recessive 1 (DSMA1)	604320	상세 확인
150	남자	-1	4	0.976197			상세 확인
68	여자	-1	5	0.973885	microcephaly 2, with or without cortical malformations	604317	상세 확인

탐색된 결과와 환자의 유사도 Plot  
**Similarity-based 2D space visualization**

Evidence 정보

Evidence ID	질환명
604320	Spinal muscular atrophy, distal, autosomal recessive 1 (DSMA1)

Search result

**탐색결과 상세보기**

입력 환자: 1  
 MRI: 2,248  
 성별: 남자

Evidence  
 ENO ID: 46  
 질환명: Spinal muscular atrophy, distal, autosomal recessive 1 (DSMA1)  
 OMIM: 604320  
 LNK: -1  
 설명:

**Input Evidence**

입력 환자 Phenotype  

HPQuem	Feature
1283	Global developmental delay
8935	Generalized neonatal hypotonia
1488	Bilateral ptosis
2020	Gastroesophageal reflux
28	Cryptorchidism
1357	Pharyngopathy
369	Overweight
11256	Motoric first degree
11354	Strabismus
2098	Respiratory distress

Evidence Phenotype  

HPQuem	Feature
1622	Premature birth
1542	Oligophrenia
1250	Generalized hypotonia
1250	Seizures

Phenotypes

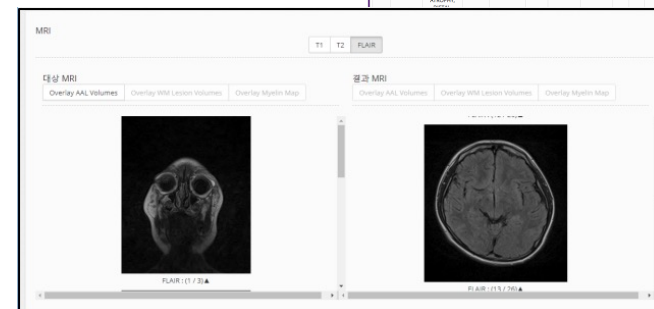
**대용 Variant**

Var ID	Gene	Disease	OMIM	Inheritance	Chr	Start	End	Ref	Alt	refRead	altRead	intRead	het_hom	Func_refGene	Ex
1	EPM2A	"CATARACT & MULTIPLE TYPES OF CTX"	116600	Autosomal dominant	chr1	16462205	16462205	A	G	34	13	47	Het	exonic	F
1	NEXT1	"XCN1 INVERSA, PARIAL, X-ACHOO"	142610	Autosomal dominant	chr1	16032753	16032753	A	G	26	17	41	Het	exonic	F
1	HNRV1	"ACTHMA, SUSCEPTIBILITY TO"	620807	Autosomal dominant, multifactorial	chr2	138771600	138771600	C	A	76	78	154	Het	exonic	F
1	GP2	"DIARRHEA MELLITUS, NONALKALIN DEPENDENT, NOEM"	128810	Autosomal dominant	chr2	157389961	157389961	C	T	57	58	107	Het	exonic	F
1	SOX4	"SEIZURES, BENIGN FAMILIAL, INADVERTENT P. SEIZ"	607740	Autosomal dominant	chr2	166221710	166221710	G	A	58	59	117	Het	exonic	F

**중요 Variant**

Var ID	Gene	Disease	OMIM	Inheritance	Chr	Start	End	Ref	Alt	refRead	altRead	intRead	het_hom	Func_refGene	DiseaseFunc_refGene
1	IGHMBP2	SPINAL MUSCULAR ATROPHY, DISTAL	604320	Autosomal recessive	chr11	68702731	68702731	C	T	2	205	207	Hom	exonic	nonstopgainof_SNV

Variants



MRI images

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%