

BERTによる情報の分類と統合データを活用した 疾患/毒性の標的探索

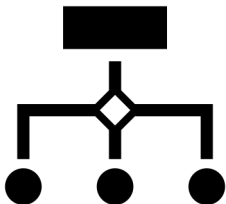
2022/9/14

緑川 淳

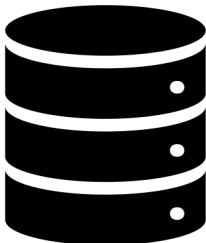
株式会社ワールドフュージョン

LSKB is scientific web

Ontology



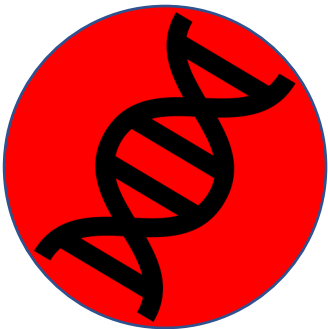
Contents



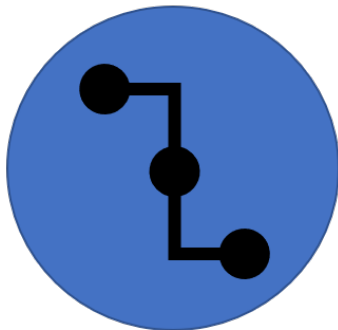
Analysis



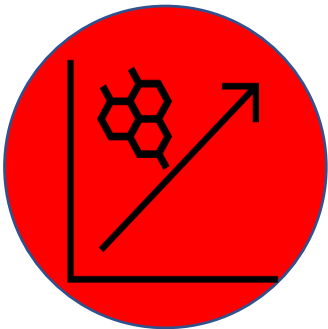
Bioinformatics



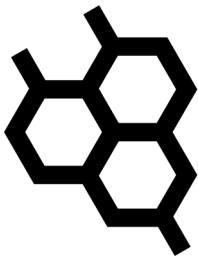
Interaction



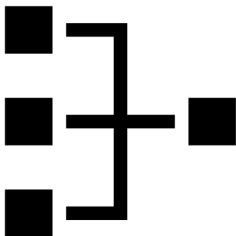
Cheminformatics



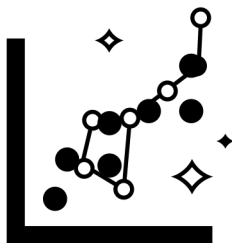
Structure



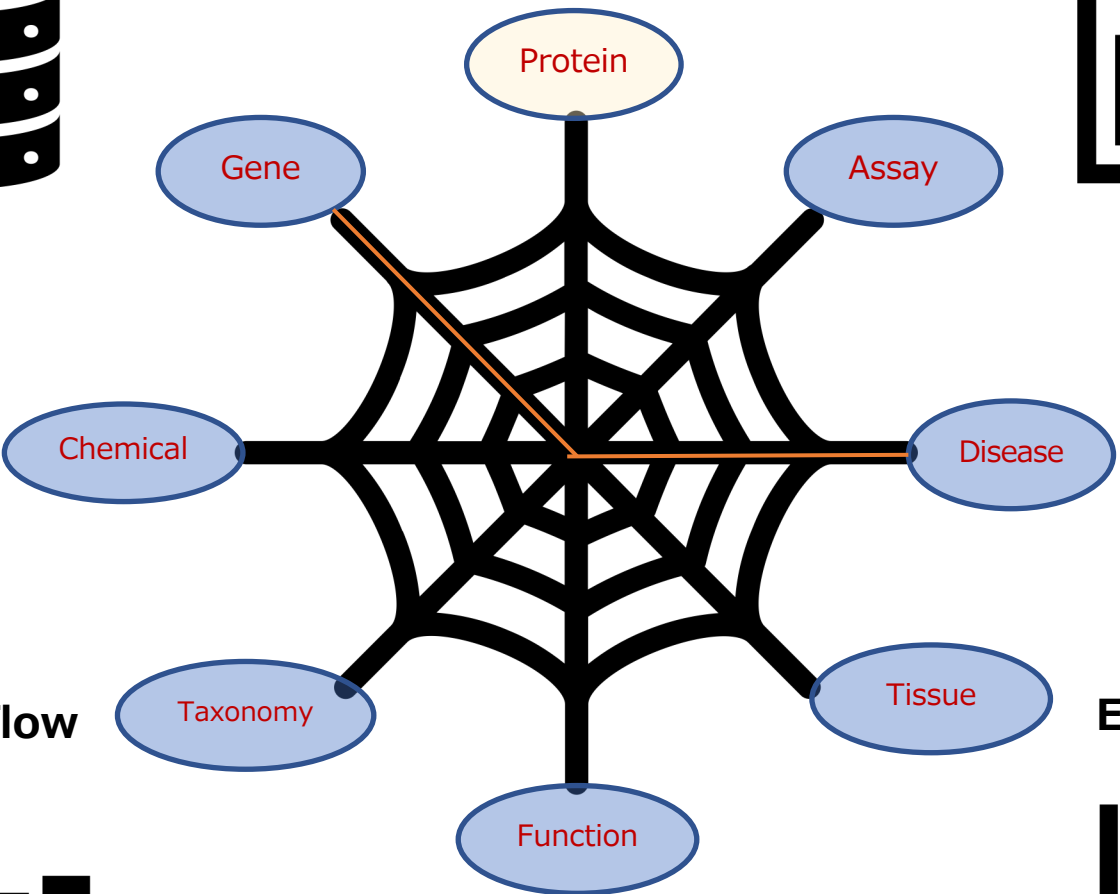
Workflow



Elpis Map

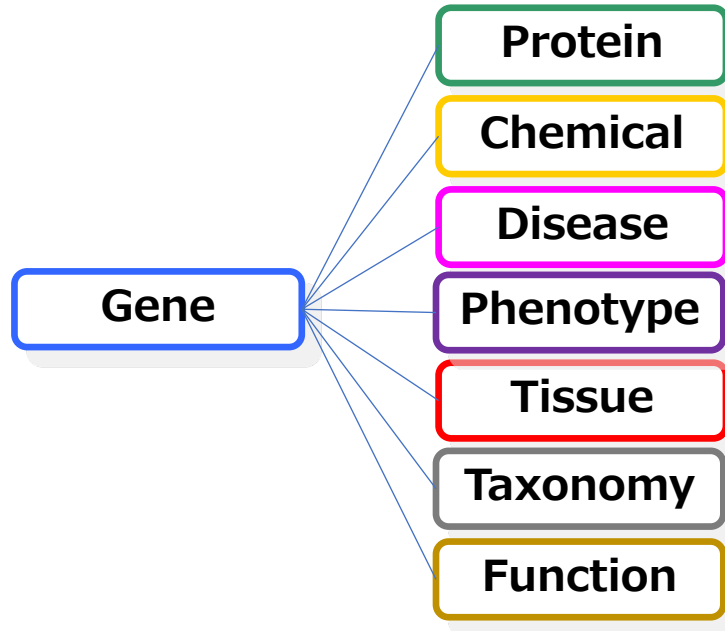
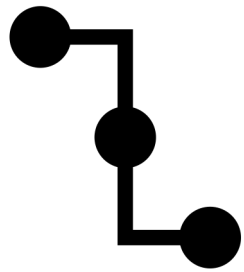


Exploration



LSKB Interaction

Interaction



1) Text-Mining

- 20 years of PubMed literature (3 levels)
- Clinical Trial
- Assay Description

2) Assay Data

- Target Gene/Protein - Chemical
- Activities Endpoint
- Mode : Inhibition agonism/antagonism
- Expression: up/down regulation
- GWAS

3) Curated Annotation

- Disease Target
- Gene Ontology
- Pathway

4) AI Curated Annotation

- Mechanism of Action
- Gene RIF
- 20 years of PubMed literature
(Toxicity associated Gene information)

BERT (Bidirectional Encoder Representations from Transformers)

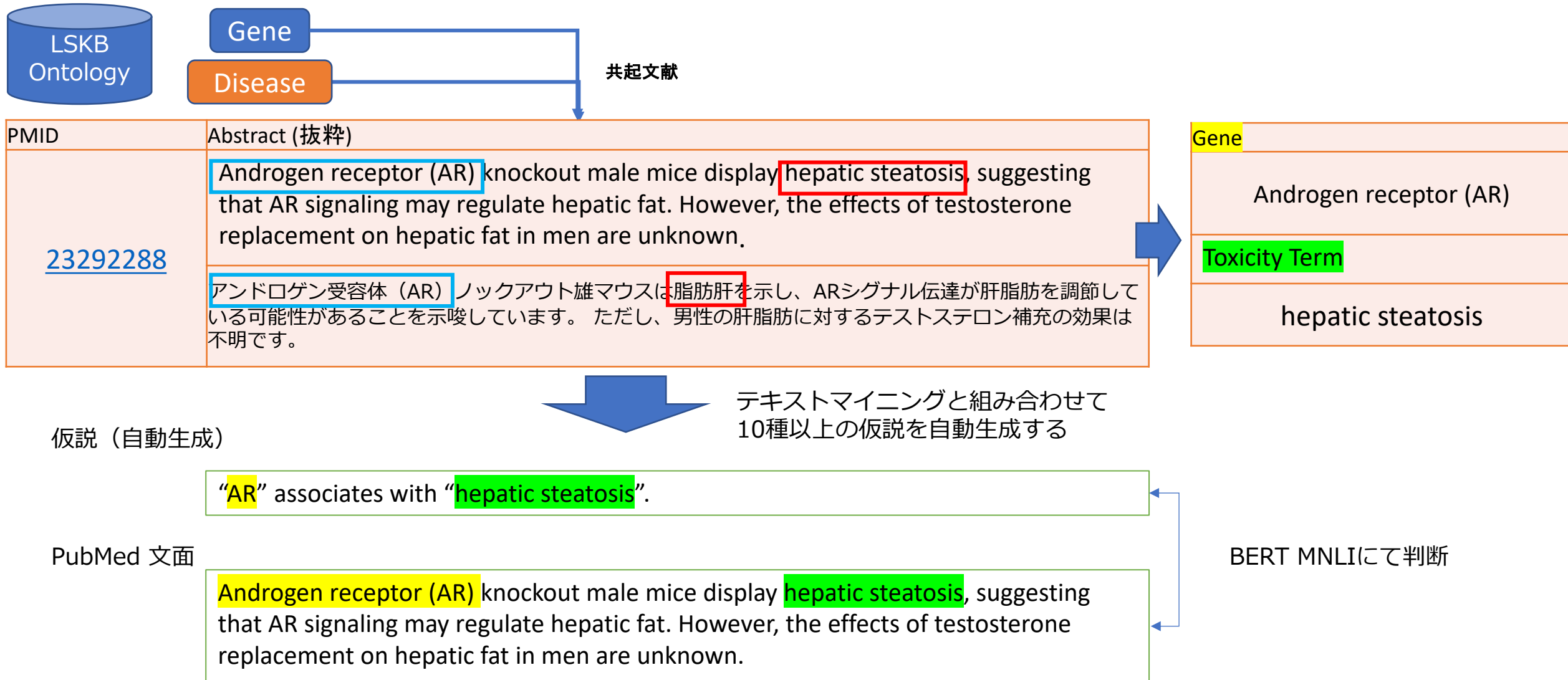
- BERTとは
 - Deep Learningにより 自然言語処理を行う手法
- BERTの特徴
 - 入力した文章の15%の単語を、確率的に別の単語で置き換えていくことで、文脈から置き換える前の単語を予測していく
 - 文章を文頭・文末の双方向から学習し、文脈を読めるようになった
 - 2018年当時の最高スコアを記録し、BERTは人工知能で初めて人間の平均の精度を超えた
- BERTのタスク
 - MNLI
 - Multi-Genre Natural Language Inferenceの略称で、GLUE（英語圏における自然言語処理の標準ベンチマーク）の指標の一つ。
 - 前提と仮説の内容上の関係に応じて「含意（entailment）」「矛盾（contradiction）」「どちらとも言えない（neutral）」といったラベルが付けられる。
- 事前学習モデル
 - **BioBERT**
 - LifeScienceに関してPubMedのAbstract、一部の本文を学習させたモデル
 - BlueBERT (NCBI)
 - PubMedBERT (Microsoft)

https://ainow.ai/2019/05/21/167211/#Next_Sentence_PredictionNSP

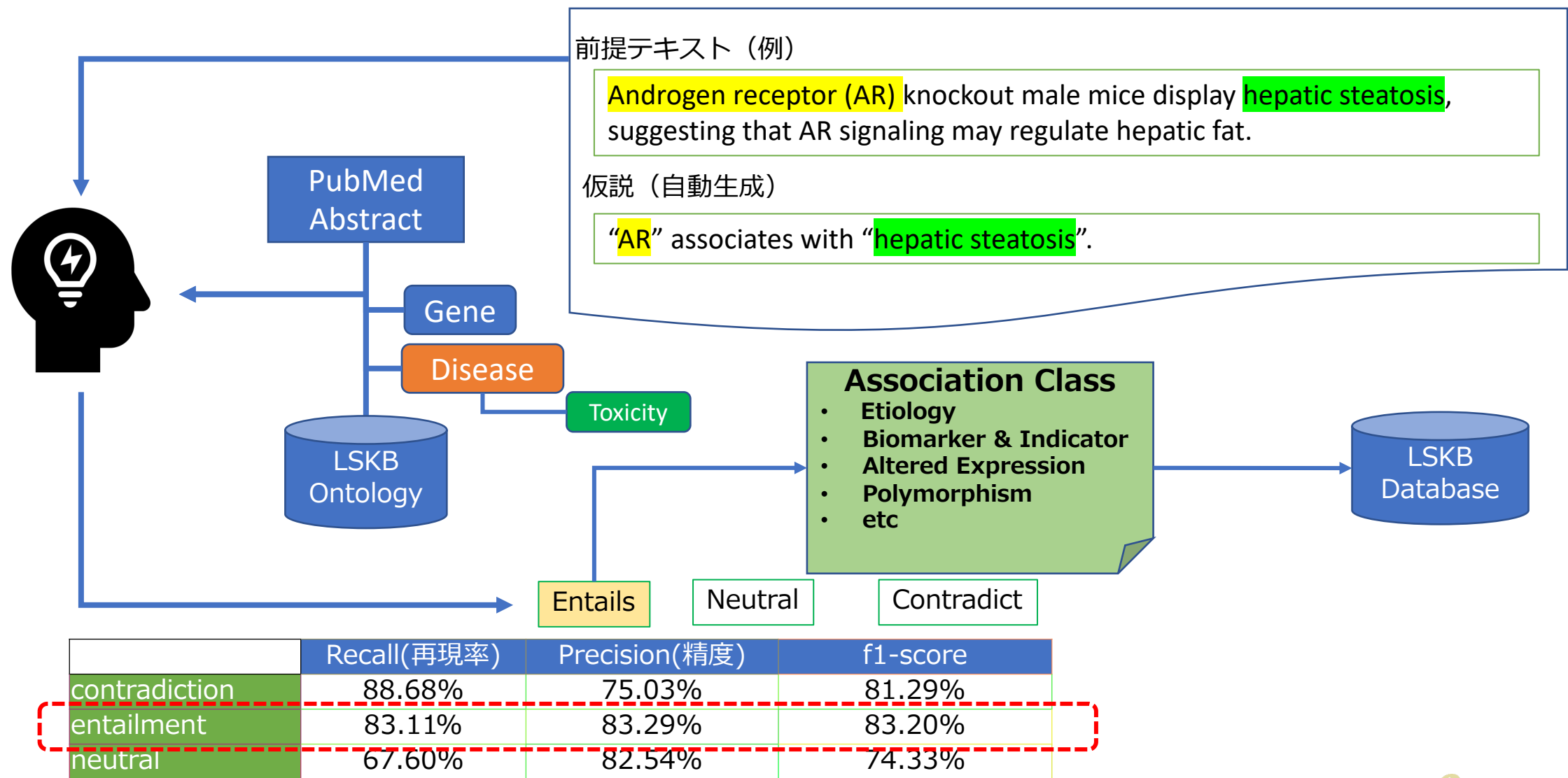
<https://buildersbox.corp->

[sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF](https://buildersbox.corp-sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF)

PubMed Abstractから Toxicity Termの認識と 仮説の生成



BERTによる遺伝子vs疾患(毒性)の関係性の判別概要



疾患関連 遺伝子情報

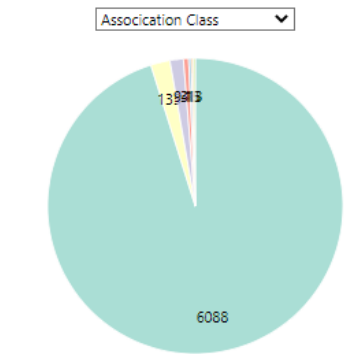
Disease Amyotrophic Lateral Sclerosis **Rare**

T047:Disease or Syndrome

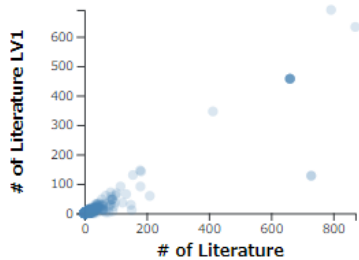
Filter Condition(s) Send Gene ID Send Gene ID to Venn Diagram A

Show 25 entries

関係性の概要分類(BERT)



X-axis : # of Literature Y-axis : # of Literature LV1



	Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	Pathway	HPO	UniProt Association	Max Phase Drug/Exp	Gene Expression	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	Gene in Disease Summary	Confidence Score
1	4744	NEFH	neurofilament heavy chain	Homo sapiens	- Altered Expression (GR) - Biomarker & Indicator (GR) - Etiology (GR) - Polymorphism (GR)	59	3	65	YES	0	0	7	6	6		0.317
2	5630	PRPH	peripherin	Homo sapiens	- Altered Expression (GR) - Etiology (GR)	35	3	65	YES	0	0	2	2	3		0.317
3	1639	DCTN1	dynactin subunit 1	Homo sapiens	Polymorphism (GR)	15	2	65	YES	0	0	3	1	9		0.3
4	23435	TARDBP	TAR DNA binding protein	Homo sapiens	- Altered Expression (GR) - Biomarker & Indicator (GR) - Etiology (GR) - Inhibition (GR) - Polymorphism (GR)	793	0	65		0	1	112	65	22		0.292

Gene: TARDBP: TAR DNA binding protein

Disease: Amyotrophic Lateral Sclerosis

Filter Condition(s) Send PMID Send PMID to Venn Diagram A

Show 25 entries

Classification

[Etiology] count: 39

PMID	PubMed Title	GeneRIF	Classification	Last Update Timestamp	Found Synonyms
32253937	A novel mutation in TARDBP segregates with amyotrophic lateral sclerosis in a large family with early onset and fast progression.	A novel mutation in TARDBP segregates with amyotrophic lateral sclerosis in a large family with early onset and fast progression.	Polymorphism	2021/06/12	Amyotrophic Lateral Sclerosis
32758693	"STRESSED OUT": The role of FUS and TDP-43 in amyotrophic lateral sclerosis.	"STRESSED OUT": The role of FUS and TDP-43 in amyotrophic lateral sclerosis.	Etiology	2021/05/08	Amyotrophic Lateral Sclerosis
			Polymorphism	2021/05/01	Amyotrophic Lateral Sclerosis
			Polymorphism	2021/01/23	Amyotrophic Lateral Sclerosis
			Polymorphism	2020/12/05	Amyotrophic Lateral Sclerosis
			Polymorphism	2020/09/12	Amyotrophic Lateral Sclerosis
			Biomarker & Indicator	2020/07/18	Amyotrophic Lateral Sclerosis
			Polymorphism	2020/07/04	Amyotrophic Lateral Sclerosis
31235914	Cryo-EM structures of four polymorphic TDP-43 amyloid cores.	Presented are the features of TDP-43 fibrils that confer both reversibility and irreversibility by determining structures of two segments reported to be the pathogenic cores of human TDP-43 aggregation: SegA (residues 311-360), which forms three polymorphs, all with dagger-shaped folds; and SegB A315E (residues 288-331 containing the amyotrophic lateral sclerosis hereditary mutation A315E), which forms R-shaped folds.	Polymorphism	2020/02/29	Amyotrophic Lateral Sclerosis
29562314	TDP-43 regulates the alternative splicing of hnRNP A1 to yield an aggregation-prone isoform.	amyotrophic lateral sclerosis patients with documented TDP-43 pathology showed neuronal hnRNP A1B cytoplasmic accumulation, indicating that TDP-43 regulates the alternative splicing of hnRNP A1 to yield an aggregation-prone isoform.	Etiology	2019/07/13	Amyotrophic Lateral Sclerosis

GeneRIF

A novel mutation in TARDBP segregates with amyotrophic lateral sclerosis in a large family with early onset and fast progression.

Classification

- Polymorphism

Target Explorer : Glaucoma related Diseases



Input :

- Angle Closure Glaucoma
- Glaucoma, Open-Angle
- Hydrophthalmos
- Low Tension Glaucoma

Filter Condition(s) Send Gene ID Send Gene ID to Venn Diagram A

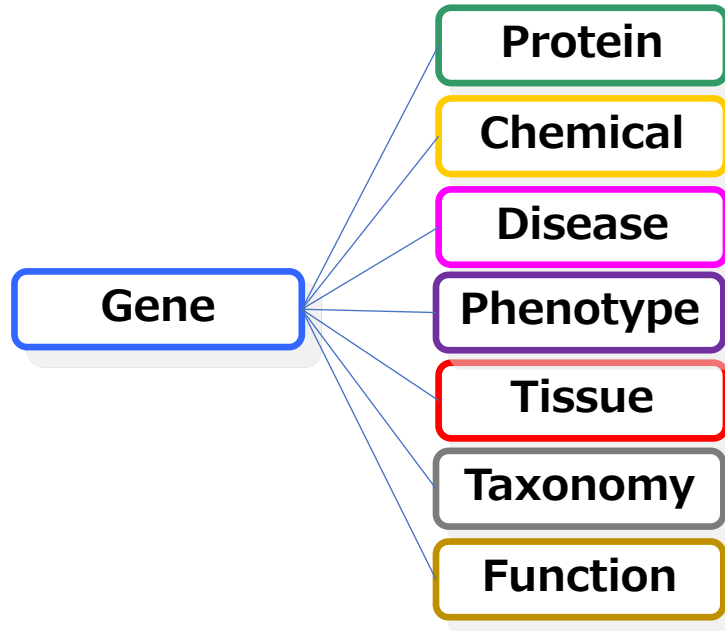
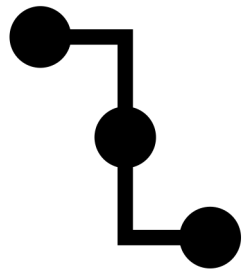
Show 25 entries

Column visibility Copy Excel

	Rank	Gene ID	Gene Symbol	Gene Title	Organism	# of Disease	Association Class	# of OMIM	# of MedGen	HPO	UniProt Association	Gene Expression	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	# of Literature	Pathway	Max Phase Drug/Exp	Confidence Score Mean	Confidence Score Max
1	1	5737	 PTGFR	prostaglandin F receptor	Homo sapiens	9	- Etiology (GR) - Polymorphism (GR) - Biomarker & indicator (PM) - Etiology (PM) - Polymorphism (PM)	0	0	0		0	3	3	0.17	18	0	4	0.24	
2	2	154	 ADRB2	adrenoceptor beta 2	Homo sapiens	9	- Etiology (GR) - Polymorphism (GR) - Biomarker & indicator (PM) - Etiology (PM) - Polymorphism (PM)	0	0	0		0	2	1	0.03	2	0	4	0.24	
3	3	153	 ADRB1	adrenoceptor beta 1	Homo sapiens	9	- Etiology (GR) - Polymorphism (GR)	0	0	0		0	1	1	0.0	0	0	4	0.23	
4	4	1545	 CYP1B1	cytochrome P450 family 1 subfamily B member 1	Homo sapiens	7	- Altered Expression (GR) - Biomarker & indicator (GR) - Enhancement (GR) - Etiology (GR) - Inhibition (GR) - Polymorphism (GR) - Altered Expression (PM) - Biomarker & indicator (PM) - Etiology (PM) - Inhibition (PM) - Polymorphism (PM)	2	0	3	YES	0	93	76	4.52	184	0	0	0.23	
5	6	760	 CA2	carbonic anhydrase 2	Homo sapiens	6	- Biomarker & indicator (PM) - Etiology (PM)	0	0	0		0	0	0	0.1	57	1	4	0.22	
6	5	4653	 MYOC	myocilin	Homo sapiens	7	- Altered Expression (GR) - Biomarker & indicator (GR) - Enhancement (GR) - Etiology (GR) - Inhibition (GR) - Polymorphism (GR)	0	0	2	YES	0	111	88	4.7	502	0	0	0.21	

LSKB Interaction

Interaction



1) Text-Mining

- 20 years of PubMed literature (3 levels)
- Clinical Trial
- Assay Description

2) Assay Data

- Target Gene/Protein - Chemical
- Activities Endpoint
- Mode : Inhibition agonism/antagonism
- Expression: up/down regulation
- GWAS

3) Curated Annotation

- Disease Target
- Gene Ontology
- Pathway

4) AI Curated Annotation

- Mechanism of Action
- Gene RIF

5) Misc

- Protein Pocket Similarity

Protein A



Protein B

Summary

PDB [11]

Annotation

Similar Protein [353]

Drug [9]

Disease [125]

Assay

keyword : "cdk1_human"

CDK1_HUMAN

> UniProt (SwissProt)

> Alpha Fold

Protein Names	Cyclin-dependent kinase 1 (2.7.11.22) (2.7.11.23) (Cell division control protein 2 homolog) (Cell division protein kinase 1) (p34 protein kinase)
Taxonomy	<i>Homo sapiens</i> [Taxonomy ID : 9606] > NCBI

→ Filter Condition(s) Send UniProt EntryName Send UniProt Entry Name to Venn Diagram A ▾

Protein	UniProt EntryName	Known binding sites	Putative binding sites	BLAST	DELTA-BLAST	# of Active Compound
1 Cyclin-dependent kinase 2 (2.7.11.22) (Cell division protein kinase 2) (p33 protein kinase)	CDK2_HUMAN	303	305	YES	YES	429
2 Serine/threonine-protein kinase Chk1 (2.7.11.1) (CHK1 checkpoint homolog) (Cell cycle checkpoint kinase) (Checkpoint kinase-1)	CHK1_HUMAN	59	4		YES	47
3 Epidermal growth factor receptor (2.7.10.1) (Proto-oncogene c-ErbB-1) (Receptor tyrosine-protein kinase erbB-1)	EGFR_HUMAN					
4 Glycogen synthase kinase-3 beta (2.7.11.26) (2.7.11.1) (Serine/threonine-protein kinase GSK3B)	GSK3B_HUMAN					
5 Mitogen-activated protein kinase 10 (2.7.11.24) (MAP kinase p49 3F12) (Stress-activated protein kinase 1b) (Stress-activated protein kinase INK3) (c-	MK10_HUMAN					

Protein B

Protein Similarity :

Pocket Similarity : PoSSuM

Known Binding Site:

Kpair

Putative Binding Site:

Ppair

配列ベース

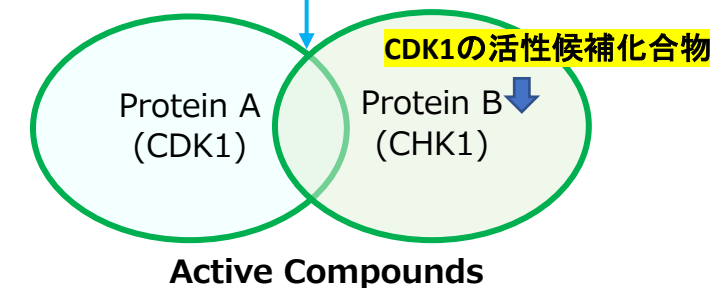
BLAST

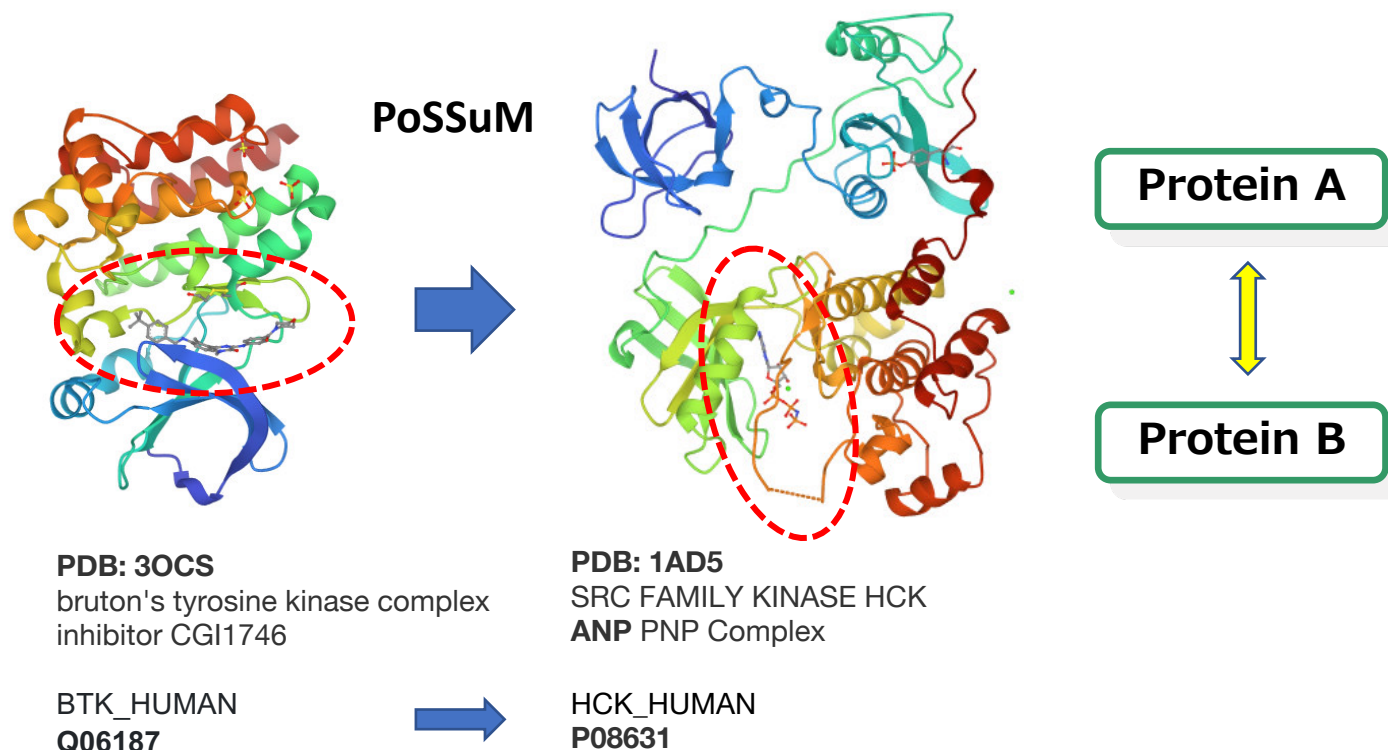
RPS-Blast

Delta-Blast

CD-Hit

共通活性化合物

PoSSuM: <http://possum.cbrc.jp/PoSSuM/>



既知ポケットの形状類似性
(Kpair : PoSSuM)

PoSSuM: <http://possum.cbrc.jp/PoSSuM/>

共通する活性化合物 :

LSKBの統合アッセイにおいて、複数のターゲットに活性のある化合物のターゲットペア

総化合物数 : 149,153

Kpair のペア数 : 322,749

疾患標的の共通性 :

同一疾患の薬剤ターゲットにおける
ポケット形状の類似タンパク質ペア

適応において複数のDrugが存在する場合、
そのうち、Phase4/3の Drugのターゲットタンパク質を比較、Kpairの類似タンパク質ペアは 45,181件

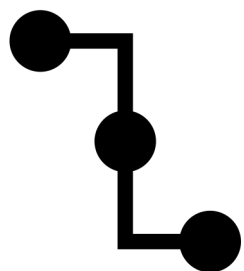
多様な疾患クラスに存在
異なるタンパク質クラスのペアに存在

活用場面

活性候補化合物の取得
Drug Repurposing
OFF-Target

LSKB Interaction

Interaction



Gene

Protein

Chemical

Disease

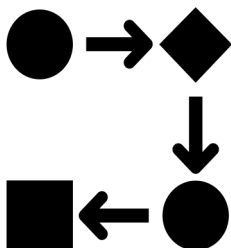
Phenotype

Tissue

Taxonomy

Function

GO-MoA database



1) Text-Mining

- 20 years of PubMed literature (3 levels)
- Clinical Trial
- Assay Description

2) Assay Data

- Target Gene/Protein - Chemical
- Activities Endpoint
- Mode : Inhibition agonism/antagonism
- Expression: up/down regulation
- GWAS

3) Curated Annotation

- Disease Target
- Gene Ontology
- Pathway

4) AI Curated Annotation

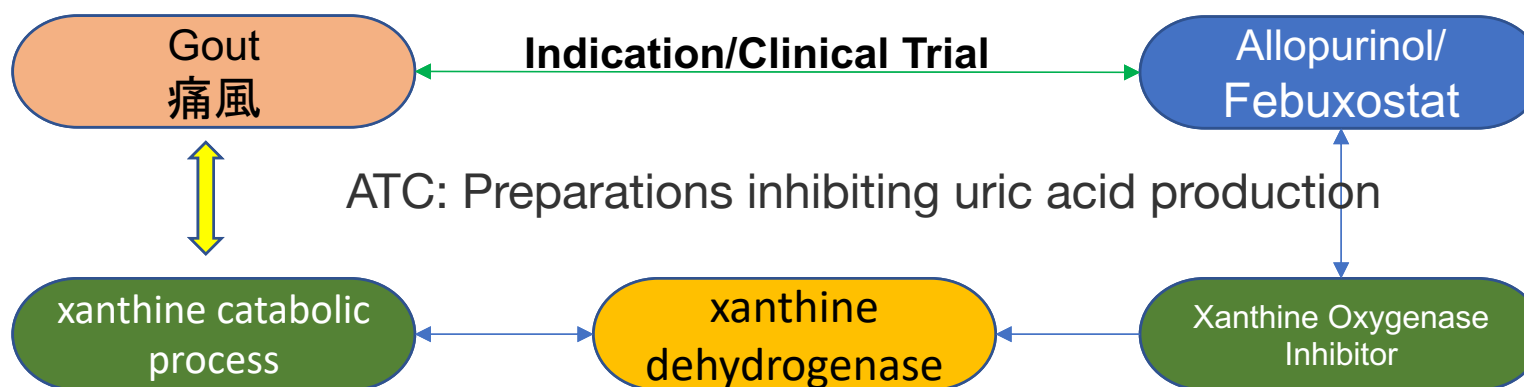
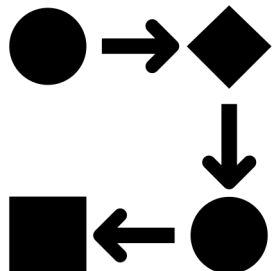
- Mechanism of Action
- Gene RIF

5) Misc

- Protein Pocket Similarity
- **GO-MoA database**

Disease related biological function database constructed by relationship of mechanism of action and the disease.

GO-MoA database

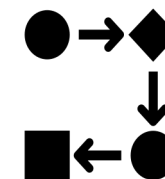


- **12,973 diseases related biological function**
- **99,610 combination of GO-MoA**
- **12,731 drugs and 2,164 targets are included.**
- **2,042 considerable human targets**
 - **functionally similar but unknown target to the disease**
- **4,019 considerable drugs**
 - **for drug repurposing**
- **750k active compounds to the targets**
In case of p-value <0.0001



Amyotrophic Lateral Sclerosis (ALS)

GO-MoA database



Disease

Amyotrophic Lateral Sclerosis Rare

T047:Disease or Syndrome

BioAssayOntology ExperimentalFactorOntology Guide to Pharmacology Medgen Orphanet UMLS

RNA

GO Top Layer :

Border of pValue :

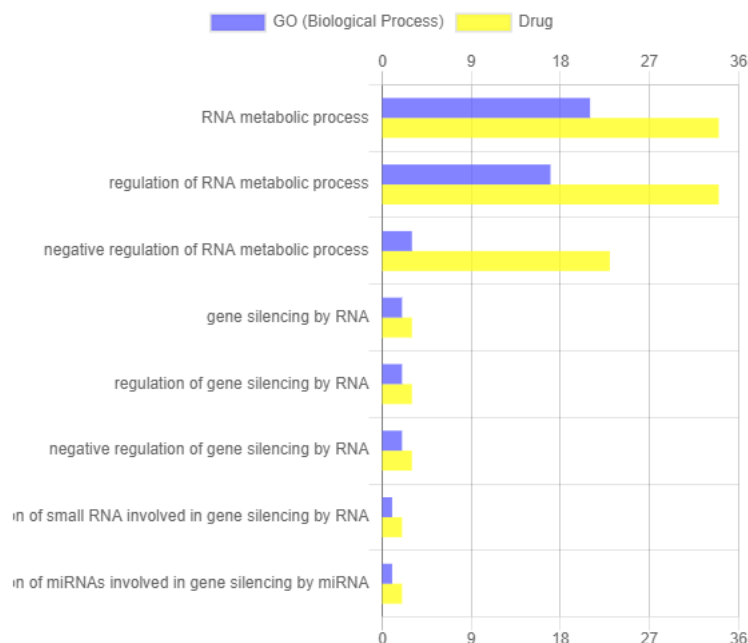
Sort by : GO

☐ 3rd ☒ 4th

1

Drug ☐

<<



Sub-categories of regulation of RNA metabolic process

☐ Gene ☒ Drug

<<

GO (Biological Process)	# of Hit Drug	# of Hit Considerable Drug	p-Value	References	# of MOA
positive regulation of transcription from RNA polymerase II promoter	28	1657	1.68E-7	0	18
negative					

Number of common drugs

	2	3	4	5	6	7	8	9
Drug	13	0	5	3	0	1	0	2
Considerable Drug	660	137	150	91	4	57	58	113

positive regulation of transcription from RNA polymerase II promoter related drugs

Send Chemical ID

Filter Data

Apply Filter(s)

Download SD File

Download TXT File

Show 25 entries

☒	Chemical	Molecule Type	Phase	Withdrawn	
☒	ropinirole	Small molecule	0		- N04 - N04 Dop
☒	Ropinirole HCl	Small molecule	0		

Showing 1 to 17 of 17 entries (filtered from 7,305 total entries)



Toxicity & Side Effect

脂肪肝に関連する遺伝子リスト

TOXICITY Hepatic steatosis

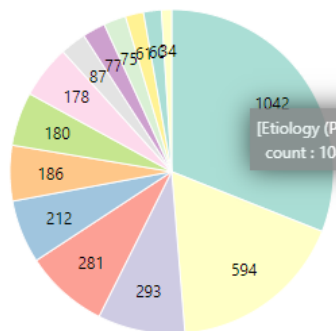
T047:Disease or Syndrome

Filter Condition(s) Send Gene ID Send Gene ID to Venn Diagram A

Show 25 entries

Column visibility Copy Excel

Association Class



X-axis: Gene ID Y-axis: LSKB_SP_ID

記載内容の分類

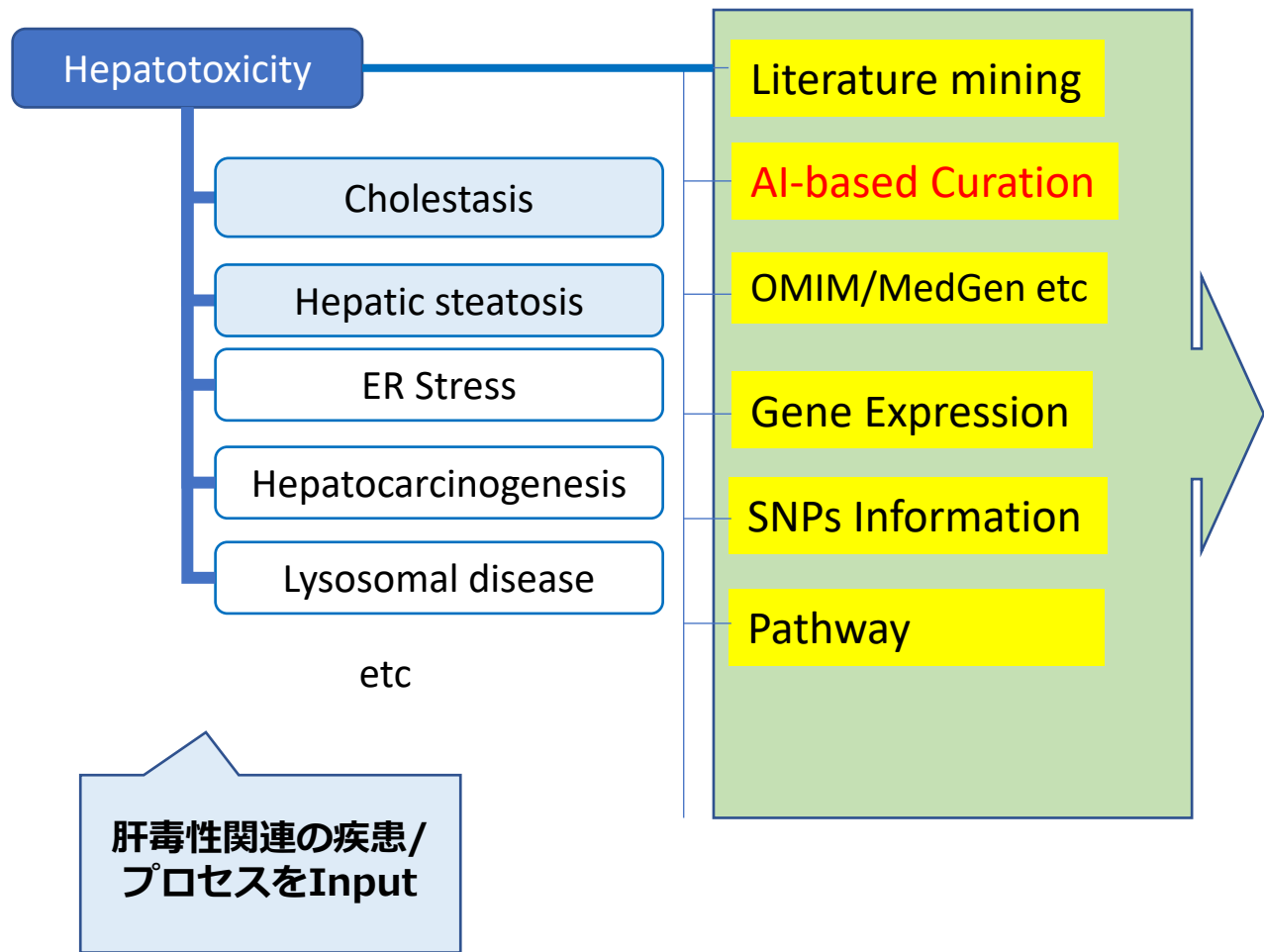
- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- etc

関係性の項目

- PubMed 文献共起
- BERTによる Association Class (PubMed/GeneRIF)
- Pathway
- Gene Expression
- GeneRIF
- SNP

	Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	Pathway	Gene Expression	Gene RIF	Gene RIF (Associations)	# of GeneRIF (Associations) for internal work	SNPs(rsID)
4	2740	GLP1R	glucagon like peptide 1 receptor	Homo sapiens	- Biomarker & indicator (PM) - Etiology (PM) - Inhibition (PM)	136	29	0	0	1	0	1	0
5	367	AR	androgen receptor	Homo sapiens	- Biomarker & indicator (PM) - Etiology (PM)	14				0	0	0	1
6	6524	SLC5A2	solute carrier family 5 member 2	Homo sapiens	- Biomarker & indicator (PM) - Etiology (PM)	37				0	0	0	0
7	3454	IFNAR1	interferon alpha and beta receptor subunit 1	Homo sapiens	- Biomarker & indicator (PM) - Down Regulation (PM) - Etiology (PM) - Polymorphism (PM)	4				0	0	0	0
8	5406	PNLIP	pancreatic lipase	Homo sapiens	- Biomarker & indicator (PM) - Etiology (PM) - Up Regulation (PM)	9				0	0	0	0
9	1268	CNR1	cannabinoid receptor 1	Homo sapiens	- Biomarker & indicator (PM) - Down Regulation (PM) - Enhancement (PM) - Etiology (PM) - Inhibition (PM) - Up Regulation (PM)	50	20	0	0	1	0	1	7
10	7124	TNF	tumor necrosis	Homo sapiens	Biomarker & indicator (GR)	375	87	13	0	23	18	20	0

Tox. Target Explorer: Hepatotoxicity

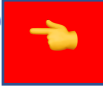


Rank	Gene ID	Gene Symbol	Gene Title	Organism	# of Disease	Association Class	HPO
						Up Regulation (PM)	
4	2	7124	[H] TNF	tumor necrosis factor	Homo sapiens	6 - Biomarker & indicator (GR) - Down Regulation (GR) - Enhancement (GR) - Etiology (GR) - Inhibition (GR) - Polymorphism (GR) - Up Regulation (GR) - Biomarker & indicator (PM) - Down Regulation (PM)	0
5	39	3309	[H] HSPA5	heat shock protein 70 class B member 5		- Inhibition (GR) - Up Regulation (GR) - Biomarker & indicator (PM) - Etiology (PM)	0
6	45	1555	[H] CYP2B6	cytochrome P450 family 2 subfamily B member 6	Homo sapiens	6 - Etiology (GR) - Polymorphism (GR) - Biomarker & indicator (PM) - Down Regulation (PM) - Enhancement (PM) - Etiology (PM) - Inhibition (PM) - Polymorphism (PM) - Up Regulation (PM)	0
7	2038	24	[H] ABCA4	ATP binding cassette subfamily A member 4	Homo sapiens	3 - Biomarker & indicator (PM) - Etiology (PM) - Polymorphism (PM)	0
8	66	1401	[H] CRP	C-reactive protein	Homo sapiens	6 - Enhancement (GR) - Etiology (GR) - Up Regulation (GR) - Biomarker & indicator (PM) - Enhancement (PM) - Etiology (PM) - Inhibition (PM) - Polymorphism (PM)	0
9	187	1649	[H] DDIT3	DNA damage inducible transcript 3	Homo sapiens	3 - Etiology (GR) - Biomarker & indicator (PM)	0

Evidenceとともに
毒性・有害事象に関わる
ターゲットをリストアップ

Evidenceとともに
毒性・有害事象に関わる
ターゲットをリストアップ

Showing 1 to 25 of 10,112 entries (filtered from 22,101 total entries)



Hepatotoxicity related Gene vs Drug

Inxight から臨床試験において急性肝炎など肝障害の報告があった79薬剤と Tox. Target Explorerにおいて上位 HepatoTox 関連 268 Gene に相当する Protein に対して活性の有無で Matrixを作成

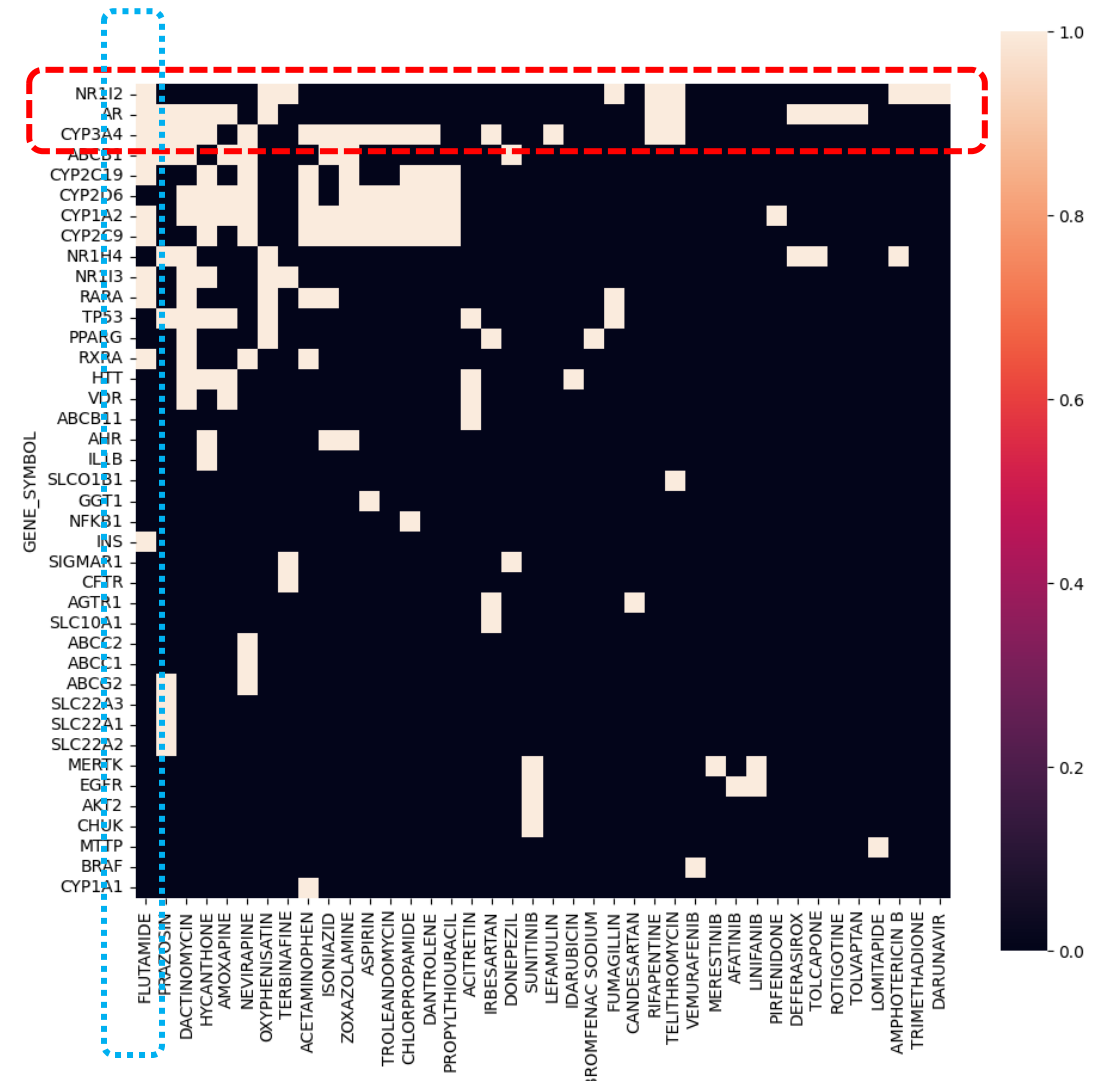
BERTにより 関連性を抽出

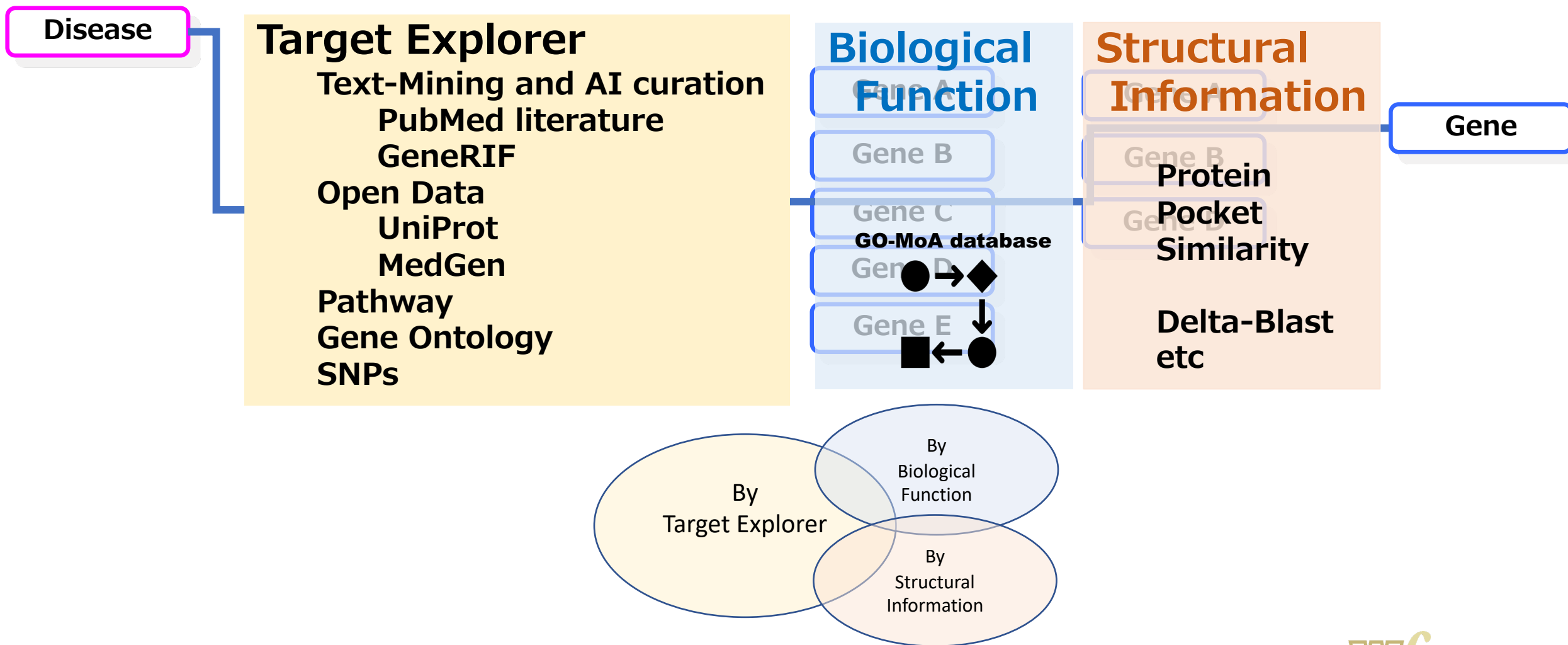
Gene: 上位の遺伝子の Evidence

- NRI12: “Pregnane X receptor”の活性化が Acetaminophen や ritonavir の hepatotoxicityを増強
- AR: アンドロゲン受容体(AR)ノックアウト雄マウスは肝脂肪症を示す
- CYP3A4: CYP3A4誘導が反応性代謝産物形成の増加を介してラパチニブ誘発肝毒性を増強

Drug:

- Flutamide
 - FDAの DILI Rankで vMost-DILI-Concern に該当
 - NRI12 : Pregnane X receptor Activator
 - CYP3A4: Inhibition
 - ABCB1 : Active





まとめ

- LSKB (Life Science Knowledge Bank) の概要
 - 多彩なInteraction
- Deep Learning (BERT)により PubMed/GeneRIF を解釈
 - 多様な仮説の自動生成により 遺伝子と疾患の関連性を取得
- 疾患関連遺伝子情報
 - 複数の関連疾患を纏めてターゲットを探索
- 疾患の治療薬のMoAと その生物学的機能を利用した GO-MoAと その利用例
- Protein Similarity
 - Protein Pocket Similarity
- 毒性 副作用関連遺伝子情報
 - 複数の疾患/プロセスから エビデンスに基づく関連遺伝子の探索
- 疾患あるいは毒性関連遺伝子の探索
 - 多様で効率的なフィルタリング

ご清聴ありがとうございました。

ご質問は

support@lskb.jp

までお願いいたします。



<https://www.lskb.jp/>