EpiGraphDB

A database and data mining platform for health data science

IEU Monthly Meeting 10 December 2019

http://docs.epigraphdb.org/slides/2019-12-ieu-meeting.pdf

Yi Liu, Benjamin Elsworth, Valeriia Haberland,
Pau Erola, Jie Zheng, Matt Lyon, Tom R Gaunt

Outline

http://docs.epigraphdb.org/slides/2019-12-ieu-meeting.pdf

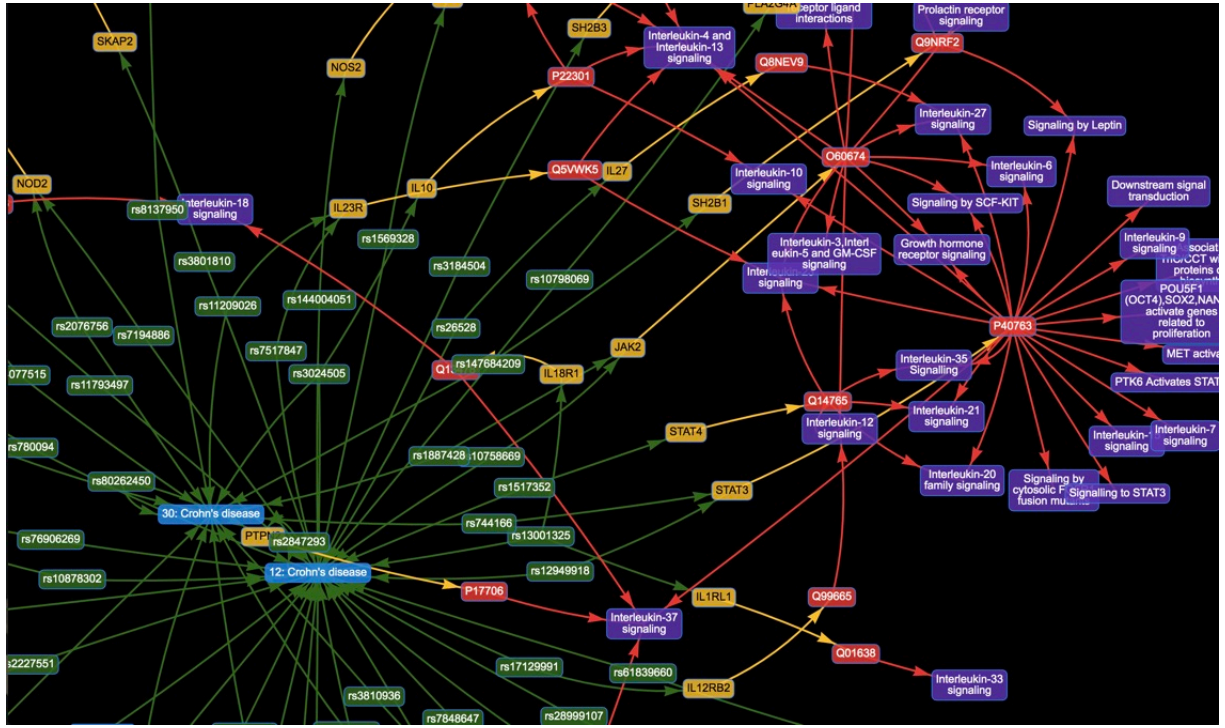
- Introduction
- EpiGraphDB project
- Use case 1: Pleiotropy
- Use case 2: Therapy response
- Use case 3: Literature
- EpiGraphDB platform
- Summary

Introduction

- Emerging trends in bioinformatics and health data science:
  - Rise of **systematic approaches** using **computational methods** in mining epidemiological relationships
  - Increasing availability of **complex, high-dimensional** epidemiological data
- EpiGraphDB as a project seeks to develop **innovative and scalable approaches** to harness their potentials to address research questions of biomedical importance.

EpiGraphDB project

- Integration of a range of data sources:
  - Systematic MR
  - Observational and genetic correlations
  - Literature-mined relationships
  - Molecular pathways
  - Protein-protein interactions
  - Drug-target relationships



- Data mining on the mechanisms of complex networks of association of risk factor / disease relationships

EpiGraphDB http://epigraphdb.org

- DB, API, web UI, R pkg, etc
- v0.2 (v0.3 in the works!)

Integrated Epidemiological Evidence

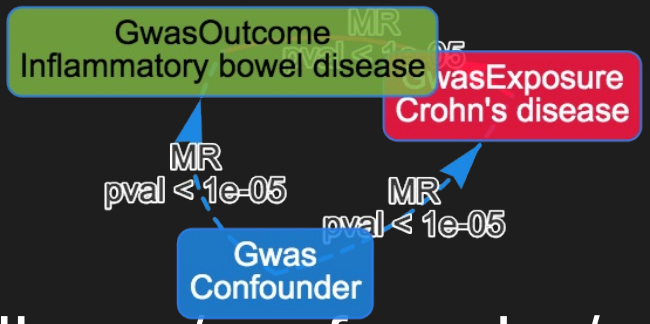
Systematic evidence from IEU studies

- IEU GWAS Database (Elsworth et al., forthcoming a);
- MR-EVE (Hemani et al., 2017)
- MELODI (Elsworth et al., 2017)
- pQTL MR (Zheng et al., 2019)
- p/eQTL MR (Zheng et al., forthcoming)
- PRS atlas (Richardson et al., 2019)
- Vectology (Elsworth et al., forthcoming b)
- Research studies by EpiGraphDB group members (http://docs.epigraphdb.org/)

External data sources

- EFO
- Gtex
- IntAct
- MeSH
- OpenTargets
- Reactome
- SemMedDB
- STRING-db
- …

Confounders

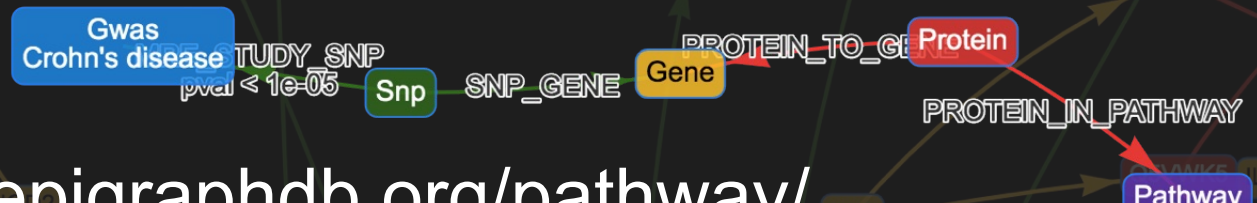


epigraphdb.org/confounder/

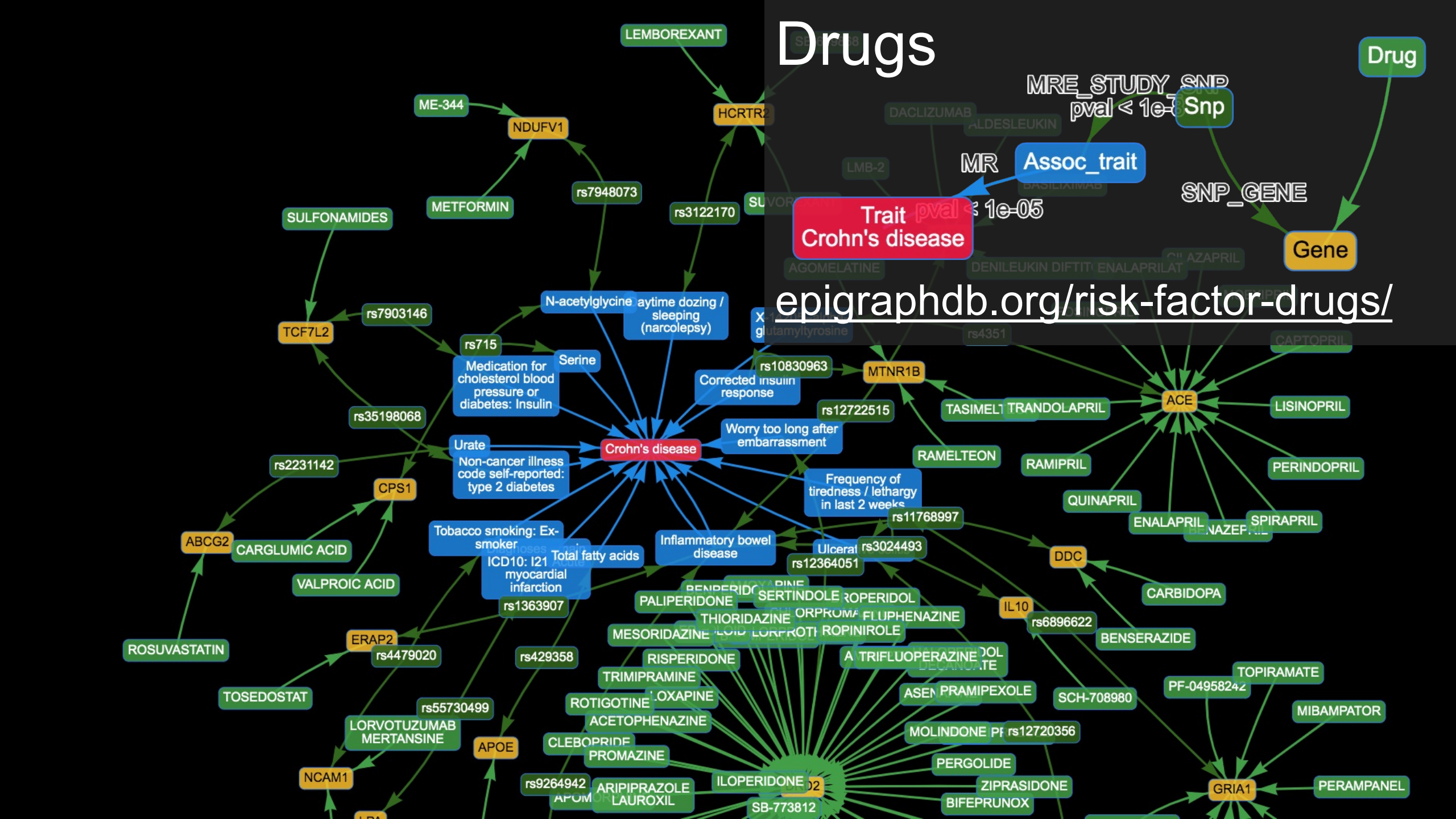
1052 Trar
855 Omega-3 fatty acids
464 Serine
996 Eczema
479 Propionylcarnitine
803 PGC cross-disorder traits
UKB-a:247 Seen a psychiatrist for nerves anxiety tension or depression
922 Concentration of small HDL particles
698 Hydroxyisovaleroyl carnitine
383 3-methyl-2-ox valerate
UKB-a:577 Diagnoses - main ICD10: N81 Female genital prolapse
Crohn's disease
294 Inflammatory
894 Lactate
497 1-linoleoylglycerophosphoethanolamine*
295 Inflammatory bowel di
860 Glutamine
9 Iron
921 Total lipids in small HDL
970 Ulcerative colitis
652 ovalerylcarnitine
845 R bisallylic total fat
UKB-a:491 Medication for cholesterol blood pressure or diabetes: Insulin
31 Inflammatory bowel disease
30 Crohn's disease
UKB-a:532 Diagnoses - main ICD10: I20 Angina pectoris
496 X-11317
32 Ulcerative colitis
720 X-14205--alpha-

883 Total lipids in large LDL
905 Cholesterol esters in medium LDL
880 Total cholesterol in large LDL
77 Waist-to-hip ratio
293 Inflammatory bowel disease
's disease
881 Cholesterol esters in large VLDL
UKB-a:71 Non-cancer illness code self-reported: cholelithiasis/gall stones
907 Concentration of medium LDL particles
Hip circumference

Pathways



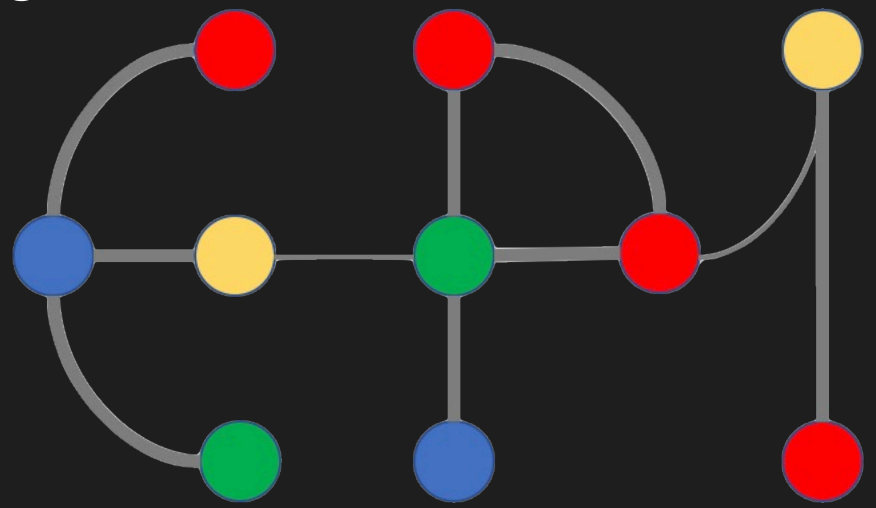
epigraphdb.org/pathway/

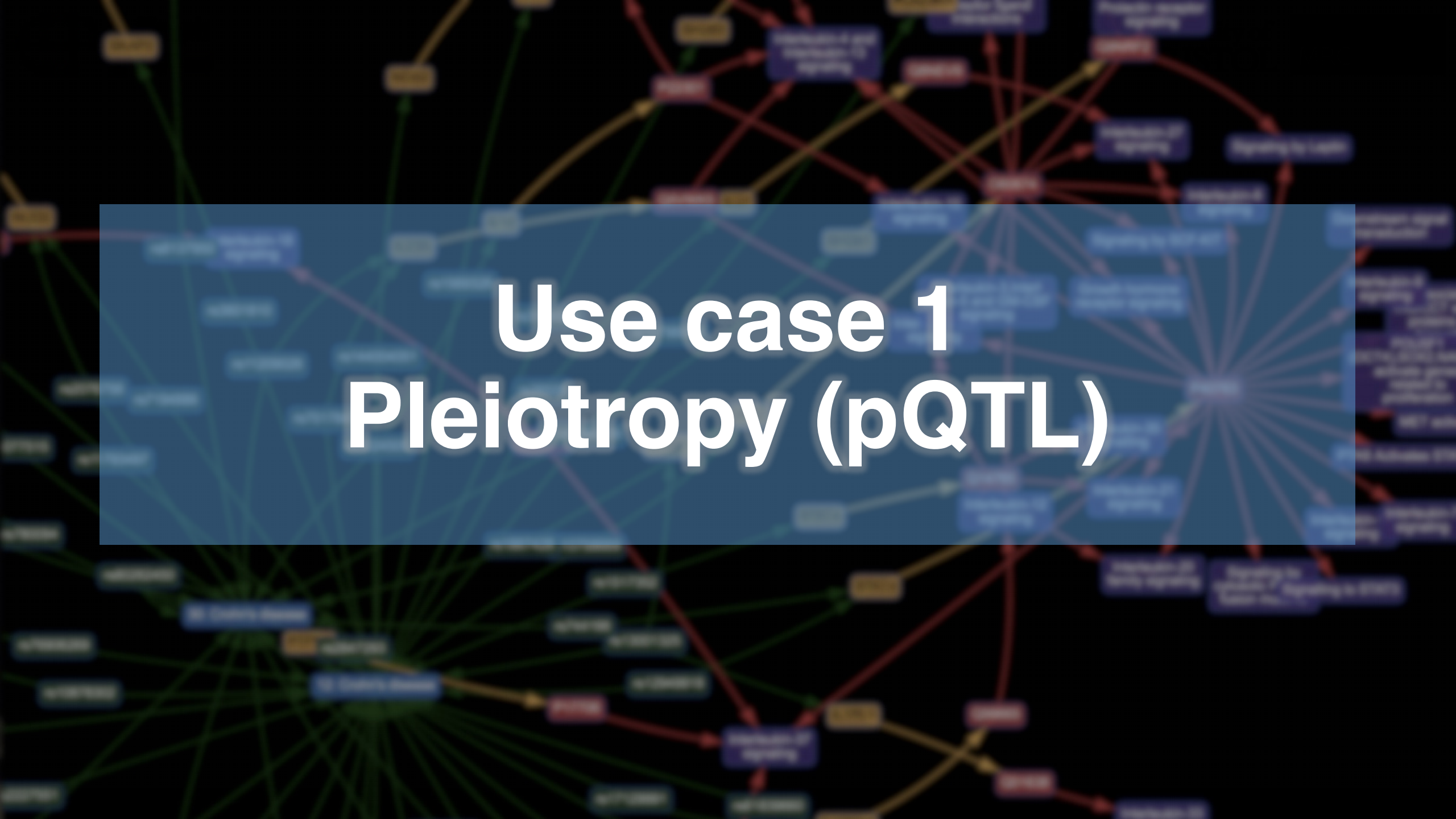

Drugs
Drug
MRE_STUDY_SNP
pval < 1e-8
Snp
MR
Assoc_trait
SNP_GENE
Trait
Crohn's disease
pval < 1e-05
Gene
epigraphdb.org/risk-factor-drugs/
LEMBOREXANT
ME-344
NDUFV1
HCRTR2
METFORMIN
rs7948073
rs3122170
SULFONAMIDES
TCF7L2
rs7903146
N-acetylglycine
Daytime dozing / sleeping (narcolepsy)
rs715
Serine
Medication for cholesterol blood pressure or diabetes: Insulin
rs35198068
Urate
Non-cancer illness code self-reported: type 2 diabetes
rs2231142
CPS1
ABCG2
CARGLUMIC ACID
VALPROIC ACID
ROSUVASTATIN
Tobacco smoking: Ex-smoker
Total fatty acids
ICD10: I21 myocardial infarction
rs1363907
ERAP2
rs4479020
TOSEDOSTAT
rs55730499
LORVOTUZUMAB MERTANSINE
NCAM1
APOE
rs429358
rs9264942
APOM
Crohn's disease
Worry too long after embarrassment
Corrected insulin response
rs10830963
MTNR1B
rs12722515
Frequency of tiredness / lethargy in last 2 weeks
Inflammatory bowel disease
rs11768997
rs3024493
rs12364051
PALIPERIDONE
SERTINDOLE
THIORIDAZINE
MESORIDAZINE
RISPERIDONE
TRIMIPRAMINE
ROTIGOTINE
ACETOPHENAZINE
PROMAZINE
ILOPERIDONE
ARIPIPRAZOLE LAUROXIL
SB-773812
TRIFLUOPERAZINE
PRAMIPEXOLE
MOLINDONE
rs12720356
PERGOLIDE
ZIPRASIDONE
BIFEPRUNOX
DDC
CARBIDOPA
IL10
rs6896622
BENSERAZIDE
SCH-708980
TOPIRAMATE
PF-04958242
MIBAMPATOR
GRIA1
PERAMPANEL
RAMELTEON
TASIMELT
TRANDOLAPRIL
ACE
LISINOPRIL
RAMIPRIL
PERINDOPRIL
QUINAPRIL
ENALAPRIL
SPIRAPRIL
CAPTOPRIL
rs4351
DACLIZUMAB
ALDESLEUKIN
LMB-2
BASILIXIMAB
AGOMELATINE
DENILEUKIN DIFTITOX
ENALAPRILAT
AZAPRIL

Integrated epidemiological evidence

http://docs.epigraphdb.org

- Causal relationships
- Association relationships
- Molecular pathways
- Literature mined / derived evidence
- Others

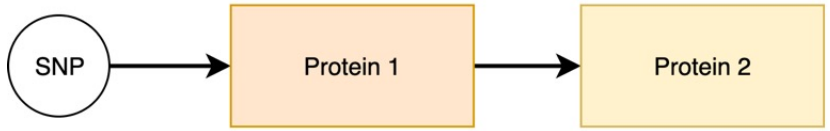



Use case 1
Pleiotropy (pQTL)

Problem (pQTL)

- Can we distinguish vertical and horizontal pleiotropic instruments using biological pathway data?

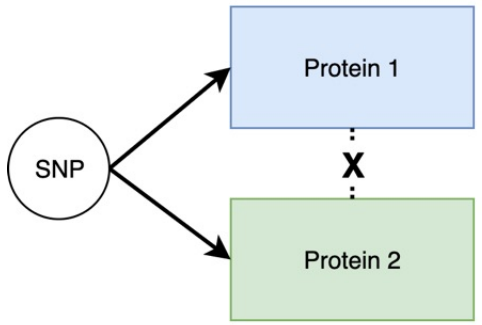
Vertical pleiotropy



SNP affect proteins on the same pathway

Valid instrument for MR

Horizontal pleiotropy

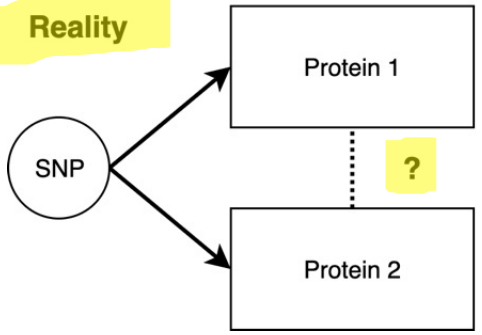


SNP associated with proteins from different pathways

Invalid instrument for MR

violates the "exclusion restriction criterion"

Reality



SNP associated with two proteins but relationship between the two proteins are missing

Valid instrument for MR?

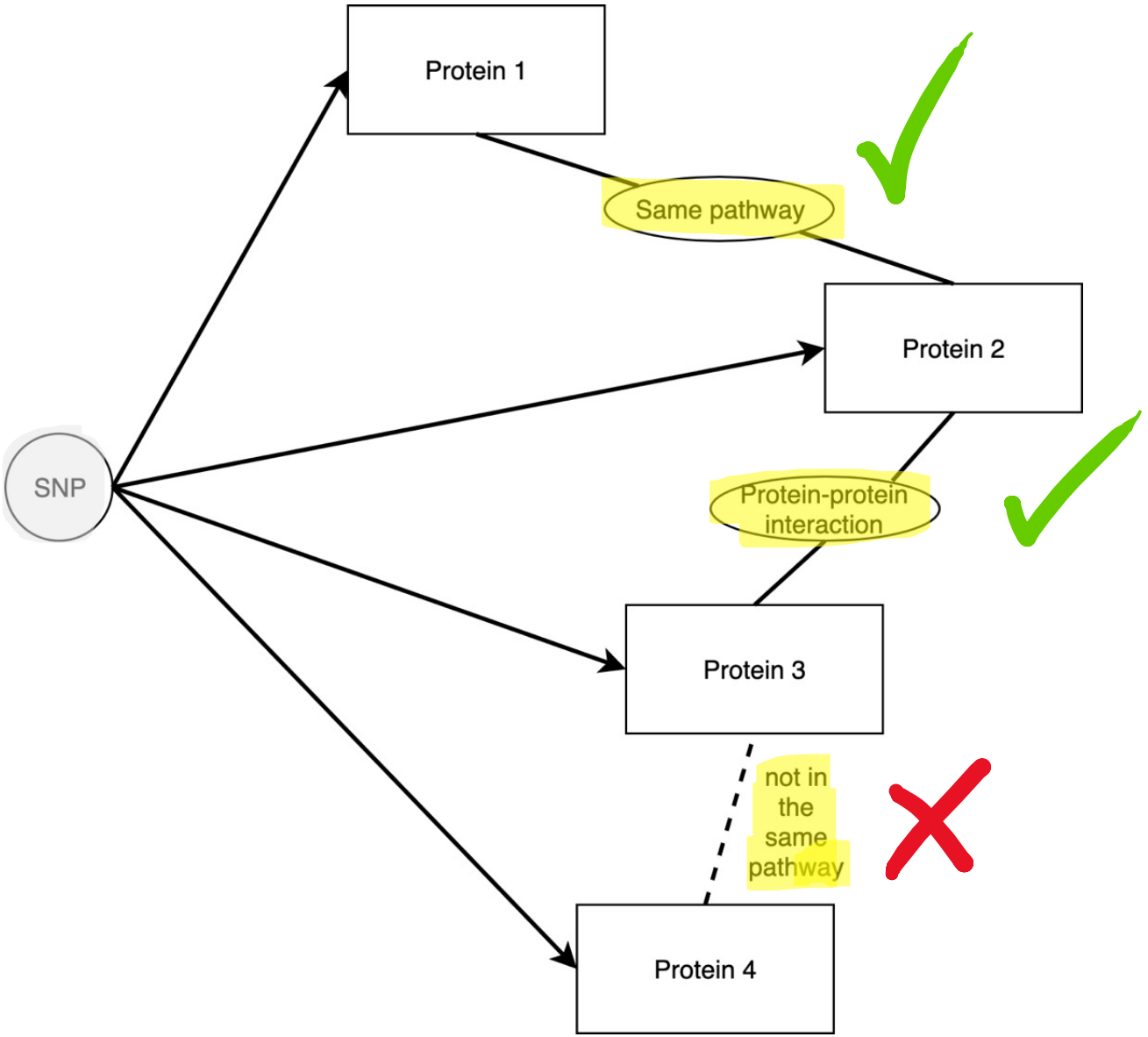
Hypothesis

For any instrument associated with multiple proteins, if

- these proteins are mapped to the same biological pathway
- exists a protein-protein interaction (PPI) between them

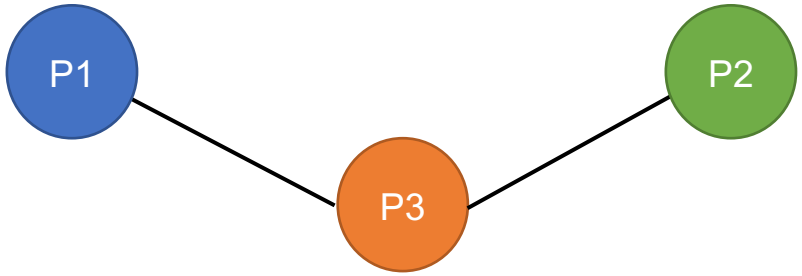
then, by definition, the instrument is more likely to act through vertical pleiotropy and it is more likely to be a valid instrument for MR.

Hypothesis

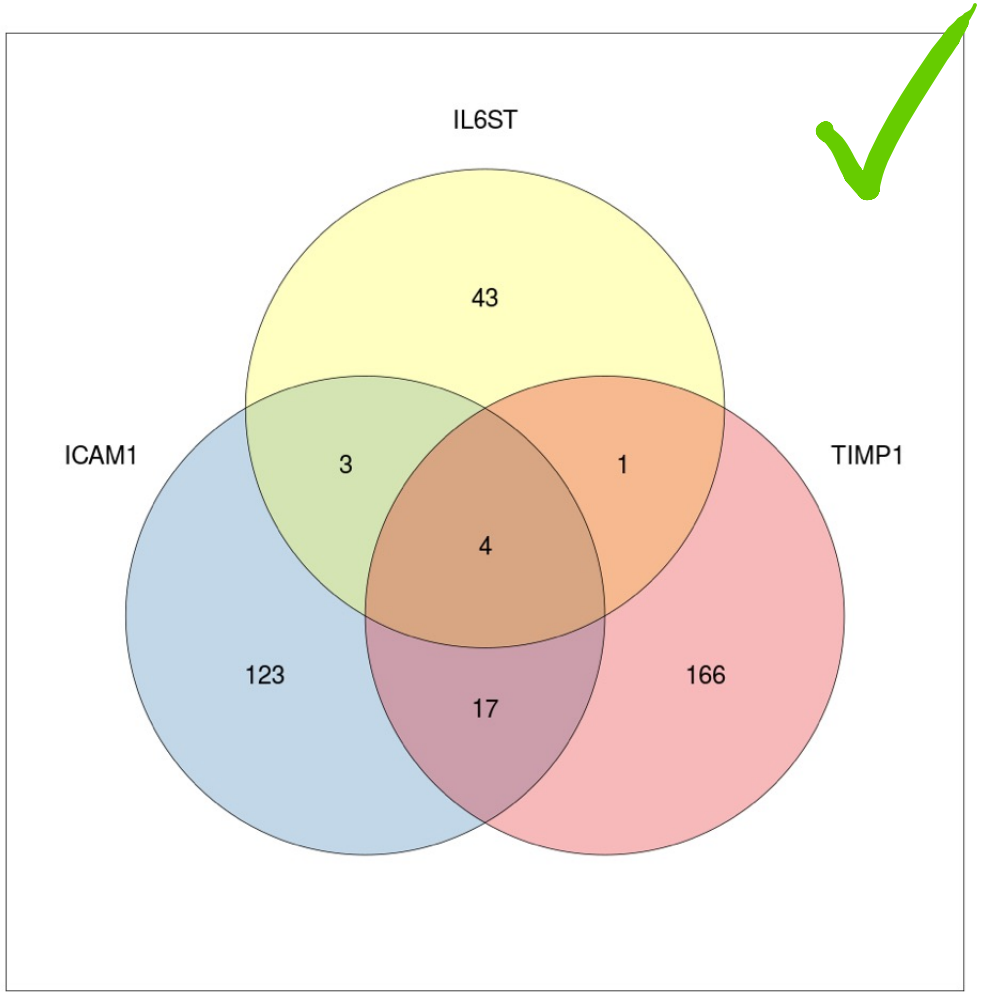

Protein 1
Same pathway
Protein 2
SNP
Protein-protein interaction
Protein 3
not in the same pathway
Protein 4

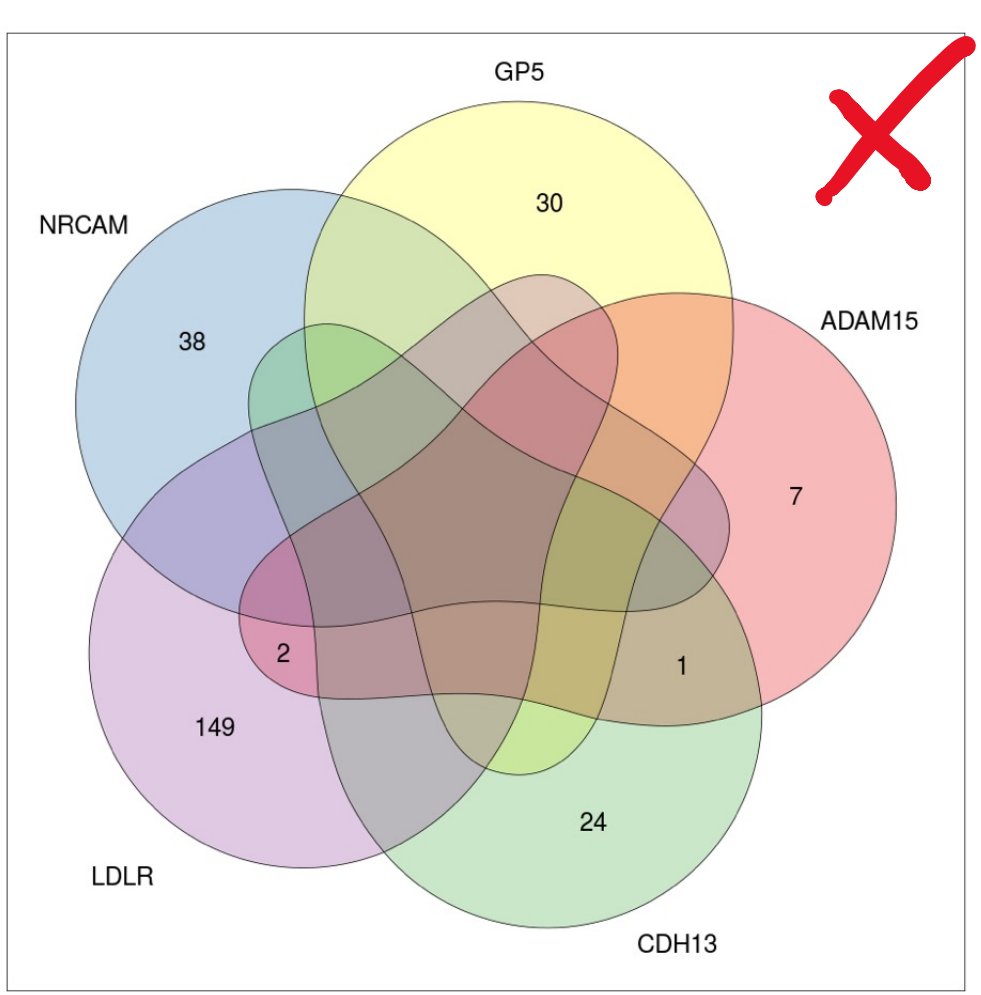
PPIs

- *We checked the number of pathways and PPIs each protein is involved in for all the instruments associated with 2 to 5 proteins*
- We used EpiGraphDB to extract high confidence PPIs from StringDB (confidence score >0.7)
  - How many PPIs they have
  - How many PPIs are shared between groups of proteins that are associated with the same SNP (or SNPs in strong LD)



PPIs - examples

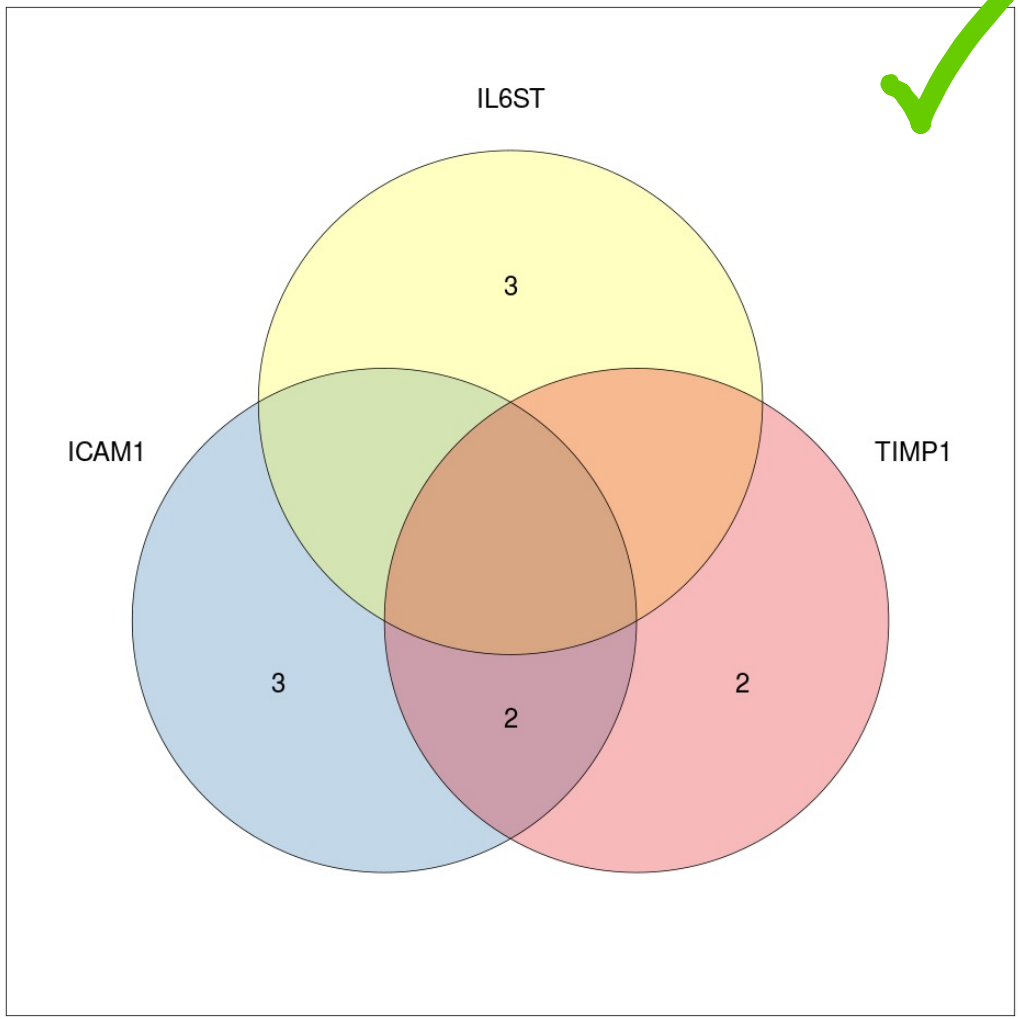


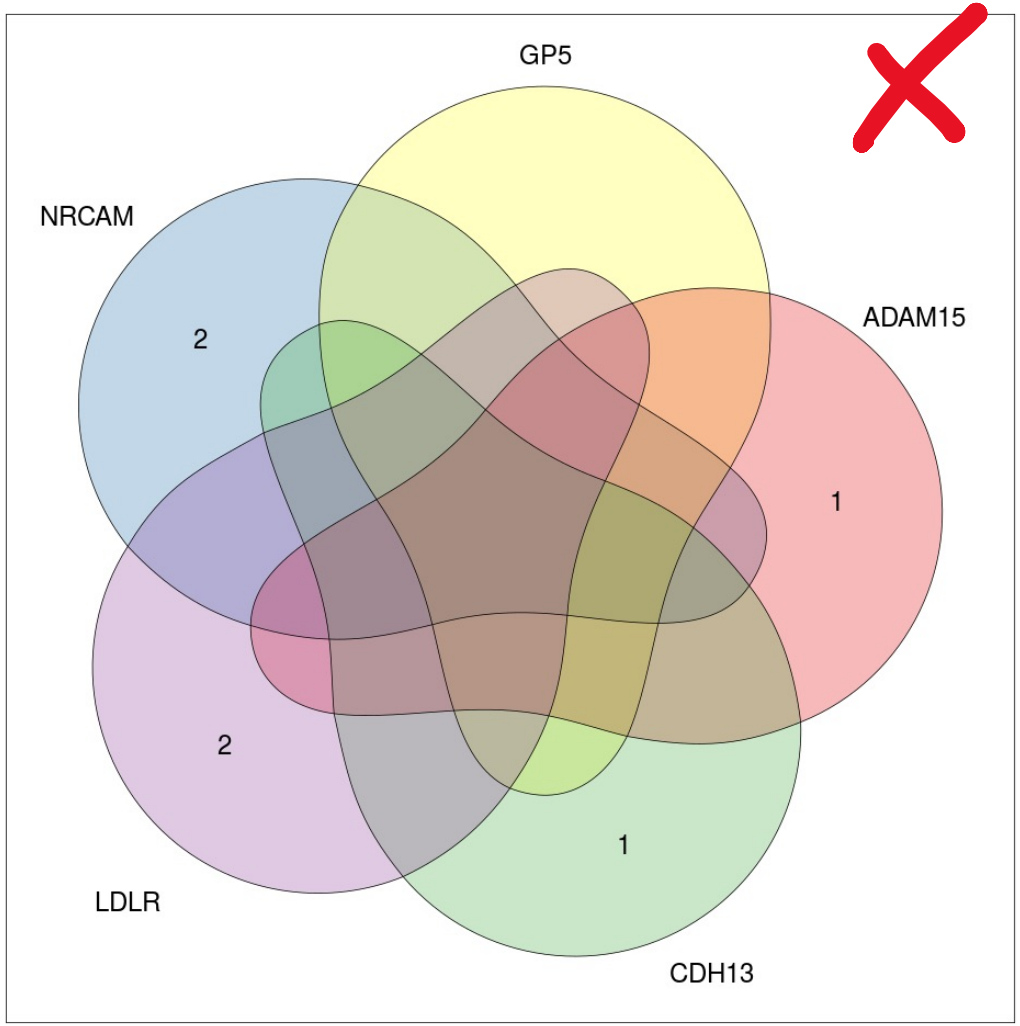


Pathways

- *We checked the number of pathways and PPIs each protein is involved in for all the instruments associated with 2 to 5 proteins*
- We used EpiGraphDB to extract pathway information from Reactome (lower level pathways)
  - Number of pathways each protein is involved in (either directly or as part of a complex)
  - How many pathways are shared between groups of proteins that are associated with the same SNP (SNPs in strong LD)

Pathways – examples





Conclusions

Jie Zheng *et al.*, Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases, *under revision*

- 263 *tier 1* instruments associated with between two and five proteins
  - Test if mapped to the same pathway or PPI
- After the analysis, 68 instruments were considered valid instruments

- **Limitation:** some pathways and PPIs that may be not included in Reactome and STRING

Summary

EpiGraphDB allows the users to evaluate the potential pleiotropic profile of genetic instruments.


Use case 2
Therapy response

Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease.

Momozawa et al.
Nat. Genetics

2011

A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.

Duerr et al.
Science

2006

IL23R polymorphisms influence phenotype and response to therapy in patients with ulcerative colitis.

Cravo et al.
Eur J Gastroenterol Hepatol.

2014

IL23R and Inflammatory Bowel Disease (IBD)

IL23R therapy response

Search for interacting* druggable** proteins:

Druggability Tier	Number of interacting proteins
Tier 1	25
Tier 2	8
Tier 3A and 3B	9

* STRING (https://string-db.org/) and IntAct (https://www.ebi.ac.uk/intact/)

** Finan et al, "The druggable genome and support for target identification and validation in drug development", *Sci. Transl. Med.* **9**, eaag1166 (2017)

Interacting proteins (Tier 1)

Gene A	Druggability Tier A	Uniprot ID A	Gene B	Druggability Tier B	Uniprot ID B
IL23R	null	Q5VWK5	IL12B	Tier 1	P29460
IL23R	null	Q5VWK5	IL12RB1	Tier 1	P42701
IL23R	null	Q5VWK5	IL23A	Tier 1	Q9NPF7
IL23R	null	Q5VWK5	JAK1	Tier 1	P23458
IL23R	null	Q5VWK5	JAK2	Tier 1	O60674
IL23R	null	Q5VWK5	STAT3	Tier 1	P40763

Summaries for IL23R Gene

Entrez Gene Summary for IL23R Gene

The protein encoded by this gene is a subunit of the receptor for IL23A/IL23. This protein pairs with the receptor molecule IL12RB1/IL12Rbeta1, and both are required for IL23A signaling. This protein associates constitutively with Janus kinase 2 (JAK2), and also binds to transcription activator STAT3 in a ligand-dependent manner. [provided by RefSeq, Jul 2008]

GeneCards Summary for IL23R Gene

IL23R (Interleukin 23 Receptor) is a Protein Coding gene. Diseases associated with IL23R include Inflammatory Bowel Disease 17 and Psoriasis 7. Among its related pathways are Cytokine Signaling in Immune system and Th17 cell differentiation. Gene Ontology (GO) annotations related to this gene include *interleukin-12 receptor binding* and *interleukin-23 receptor activity*. An important paralog of this gene is LEPR.

Alternative drug targets

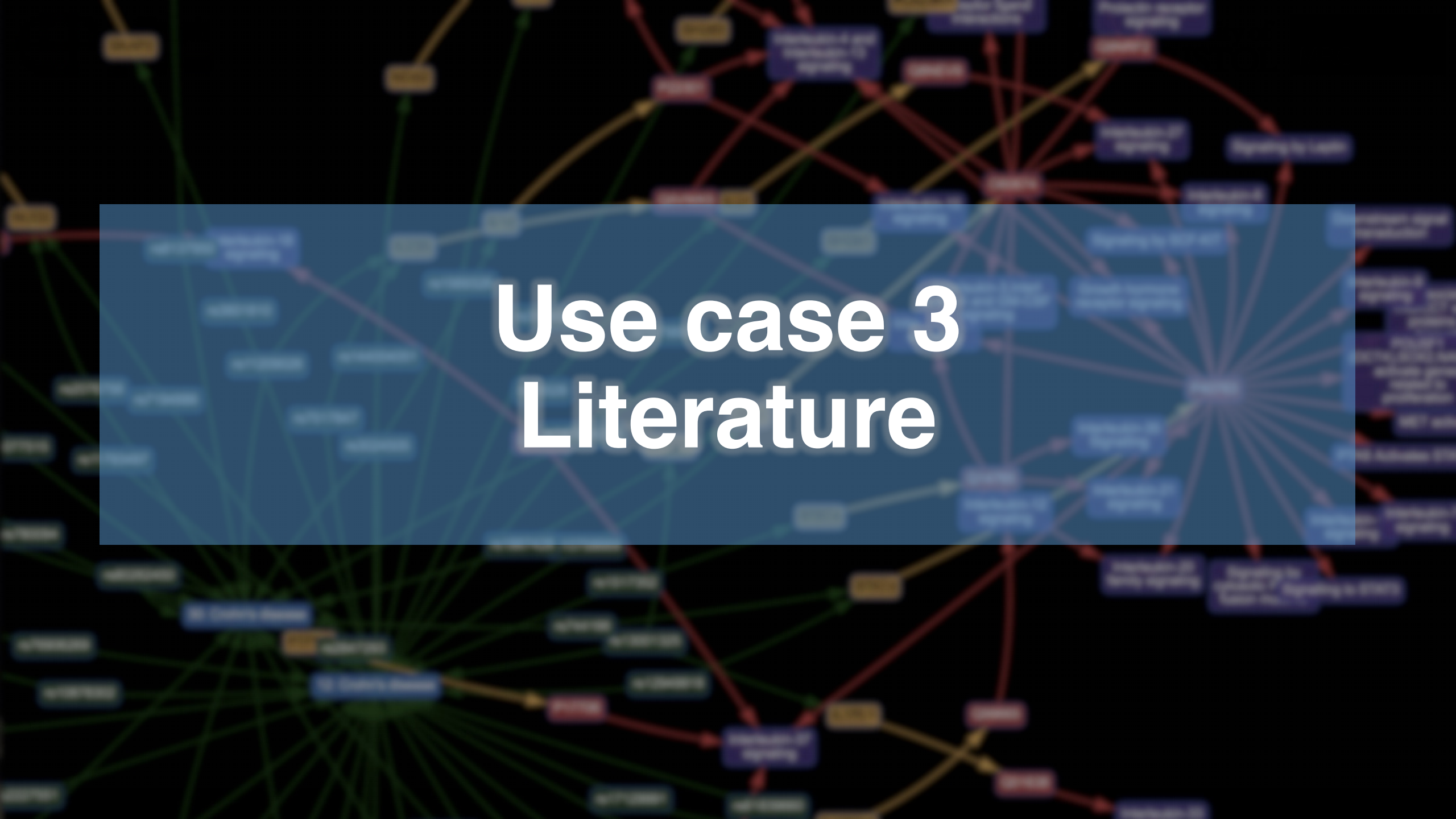
Search for MR results* for strongly related proteins (Tier 1) and their effect on IBD:

Gene	Effect size	SE	P-value	ID	Outcome
IL23R	1.5	0.0546	**2.21E-166**	294	Inflammatory bowel disease
IL12RB1	-0.0097	0.0142	0.49	294	Inflammatory bowel disease
IL12B	0.42	0.0345	**9.59E-34**	294	Inflammatory bowel disease

* Zheng, Haberland, Baird, et al. "Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases", Submitted revised version to Nat. Genetics (2019)

Summary

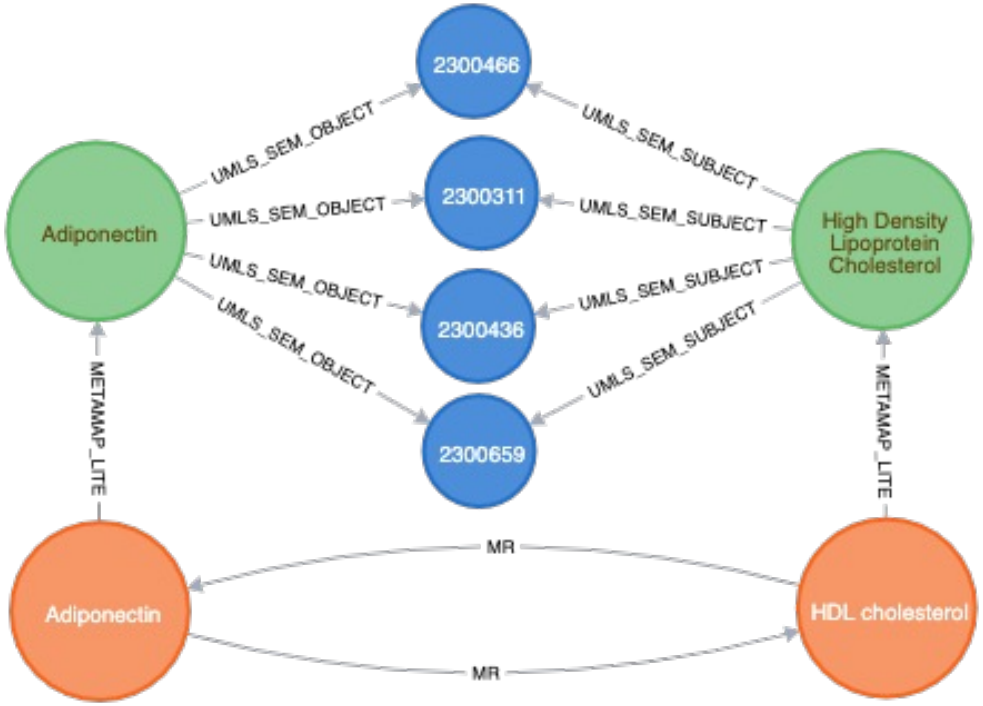
EpiGraphDB allows the users to search for therapy response related information either for the intended target or along its pathway.



Use case 3
Literature

Literature data v0.2

- Limited literature data
- Links to Publications from various places
- Links to SemMedDB via ontology matches



Literature data v0.3

v0.3 (some time next year)

- All SemMedDB and PubMed
- MELODI Lite enrichment for each GWAS

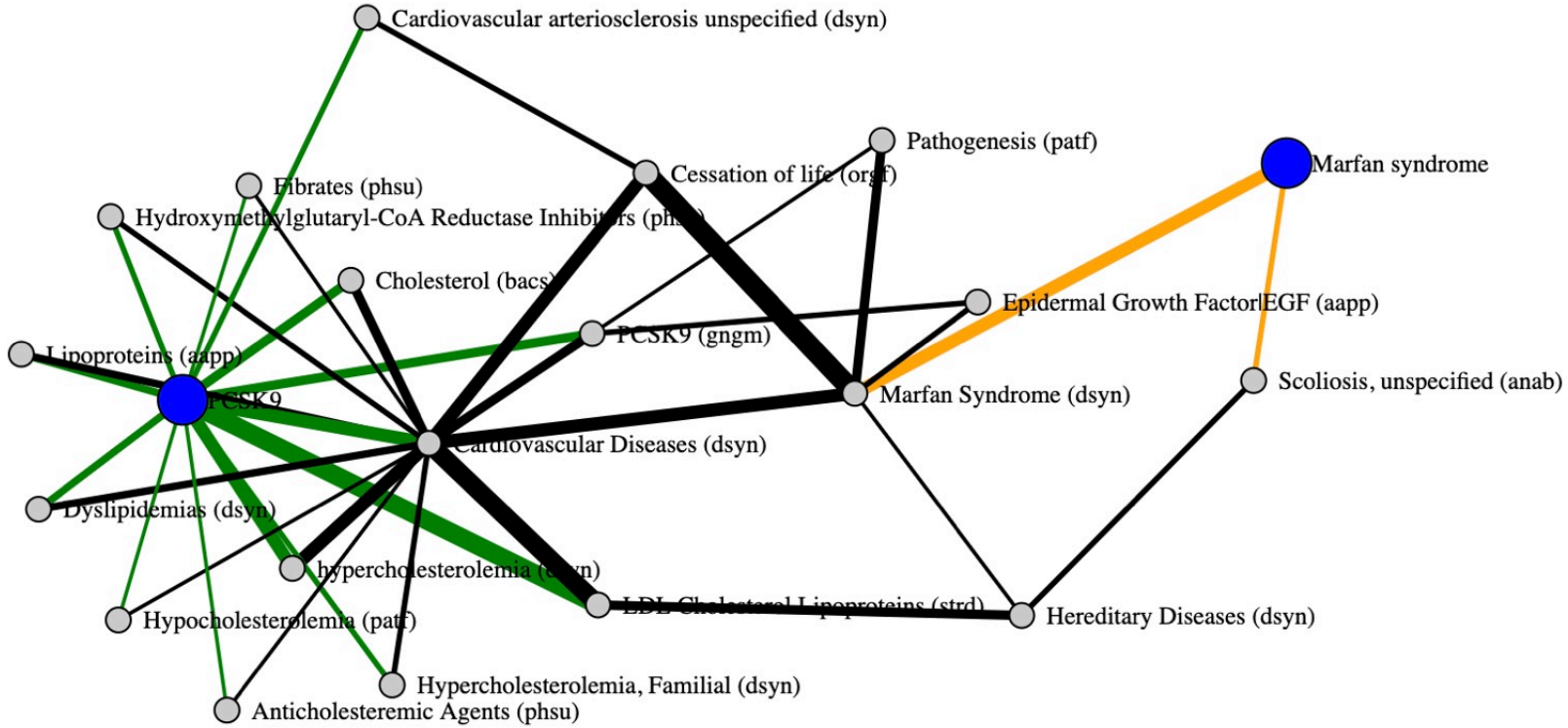
MELODI

http://melodi.biocompute.org.uk/

SemMedDB

Database of triples extracted from MEDLINE titles and abstracts, e.g.

PCSK9 (subject)
PREDISPOSES (predicate)
Cardiovascular Diseases (object)



MELODI Lite

Restricted to certain types and predicates

100x quicker

Multiple exposures and outcomes

Application programming interface:

http://textbase.biocompute.org.uk/docs/

v0.3 Literature Metrics

Genes represented in the literature	15,489 / 57,736
Pathways represented by genes in the literature	2,249 / 2,259
GWAS with literature evidence	4,226 / 11,016
Trait-MR->Trait (p<1e-20) pairs with no literature connection	1,839 / 8,830

Risk factors for Crohn's disease

1. **Find potential risk factors (no ncase value - continuous)**

```
match (g1:Gwas)-[mr:MR]->(g2:Gwas) where g2.id = 'ieu-a-30'
and not exists(g1.ncase) and mr.pval<1e-5 with order by
mr.pval asc, mr.moescore desc limit 5 with g1,g2,mr
collect(distinct(g1.id))+collect(distinct(g2.id))as g_list
```

Risk factors for Crohn's disease

1. Find potential risk factors (no ncase value - continuous)

2. **Get SNP-gene data for all GWAS**

```
match (g1:Gwas)-[mr:MR]->(g2:Gwas) where g2.id = 'ieu-a-30'
and not exists(g1.ncase) and mr.pval<1e-5 with order by
mr.pval asc, mr.moescore desc limit 5 with g1,g2,mr
collect(distinct(g1.id))+collect(distinct(g2.id))as g_list


match (gene1:Gene)<-[vg:VARIANT_TO_GENE]-(v:Variant)-
[gv:GWAS_TO_VARIANT]-(gwas:Gwas) where gwas.id in g_list and
gv.pval<1e-20 with gene1,gwas,v,vg
```

Risk factors for Crohn's disease

1. Find potential risk factors (no ncase value - continuous)
2. Get SNP-gene data for all GWAS
3. **Get literature data for all GWAS**

```
match (g1:Gwas)-[mr:MR]->(g2:Gwas) where g2.id = 'ieu-a-30'
and not exists(g1.ncase) and mr.pval<1e-5 with order by
mr.pval asc, mr.moescore desc limit 5 with g1,g2,mr
collect(distinct(g1.id))+collect(distinct(g2.id))as g_list


match (gene1:Gene)<-[vg:VARIANT_TO_GENE]-(v:Variant)-
[gv:GWAS_TO_VARIANT]-(gwas:Gwas) where gwas.id in g_list and
gv.pval<1e-20 with gene1,gwas,v,vg


optional match (gwas)-[gs:GWAS_SEM]-(s:SemmedTriple)-
[:SEM_SUB|:SEM_OBJ]-(st:SemmedTerm)-[sg:SEM_GENE]-
(gene2:Gene) where gs.pval<1e-20 with
gwas,gene1,gene2,v,s,st
```

Risk factors for Crohn's disease

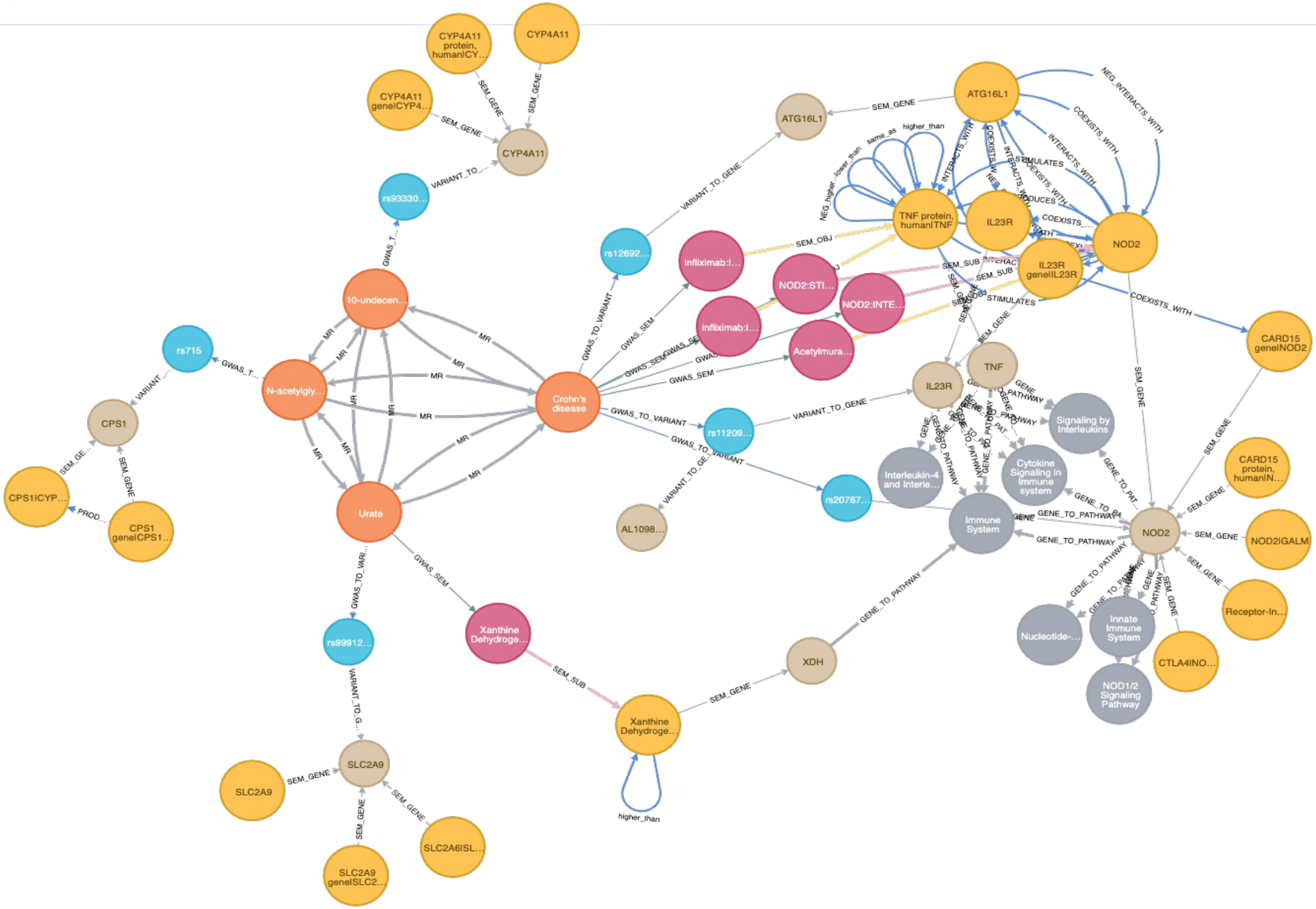
1. Find potential risk factors (no ncase value - continuous)
2. Get SNP-gene data for all GWAS
3. Get literature data for all GWAS
4. **Find shared pathways between SNP-genes and literature genes**

```
match (g1:Gwas)-[mr:MR]->(g2:Gwas) where g2.id = 'ieu-a-30'
and not exists(g1.ncase) and mr.pval<1e-5 with order by
mr.pval asc, mr.moescore desc limit 5 with g1,g2,mr
collect(distinct(g1.id))+collect(distinct(g2.id))as g_list


match (gene1:Gene)<-[vg:VARIANT_TO_GENE]-(v:Variant)-
[gv:GWAS_TO_VARIANT]-(gwas:Gwas) where gwas.id in g_list and
gv.pval<1e-20 with gene1,gwas,v,vg


optional match (gwas)-[gs:GWAS_SEM]-(s:SemmedTriple)-
[:SEM_SUB|:SEM_OBJ]-(st:SemmedTerm)-[sg:SEM_GENE]-
(gene2:Gene) where gs.pval<1e-20 with
gwas,gene1,gene2,v,s,st


optional match (gene1)-[:GENE_TO_PATHWAY]->(p:Pathway)<-
[:GENE_TO_PATHWAY]-(gene2)
```

Risk factors for Crohn's disease

1. Find potential risk factors (no ncase value - continuous)
2. Get SNP-gene data for all GWAS
3. Get literature data for all GWAS
4. Find shared pathways between SNP-genes and literature genes
5. **Map SNP-genes to literature genes**

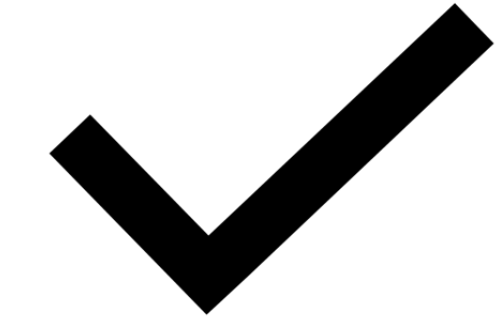
```
match (g1:Gwas)-[mr:MR]->(g2:Gwas) where g2.id = 'ieu-a-30'
and not exists(g1.ncase) and mr.pval<1e-5 with order by
mr.pval asc, mr.moescore desc limit 5 with g1,g2,mr
collect(distinct(g1.id))+collect(distinct(g2.id))as g_list


match (gene1:Gene)<-[vg:VARIANT_TO_GENE]-(v:Variant)-
[gv:GWAS_TO_VARIANT]-(gwas:Gwas) where gwas.id in g_list and
gv.pval<1e-20 with gene1,gwas,v,vg


optional match (gwas)-[gs:GWAS_SEM]-(s:SemmedTriple)-
[:SEM_SUB|:SEM_OBJ]-(st:SemmedTerm)-[sg:SEM_GENE]-
(gene2:Gene) where gs.pval<1e-20 with
gwas,gene1,gene2,v,s,st


optional match (gene1)-[:GENE_TO_PATHWAY]->(p:Pathway)<-
[:GENE_TO_PATHWAY]-(gene2)


optional match (gene1)-[:SEM_GENE]-(st2:SemmedTerm) return
gwas,gene1,gene2,v,s,st,p,st2;
```



- 6 SNPs
- 9 genes
- 6 semantic
  - 292 papers
  - 232 journals

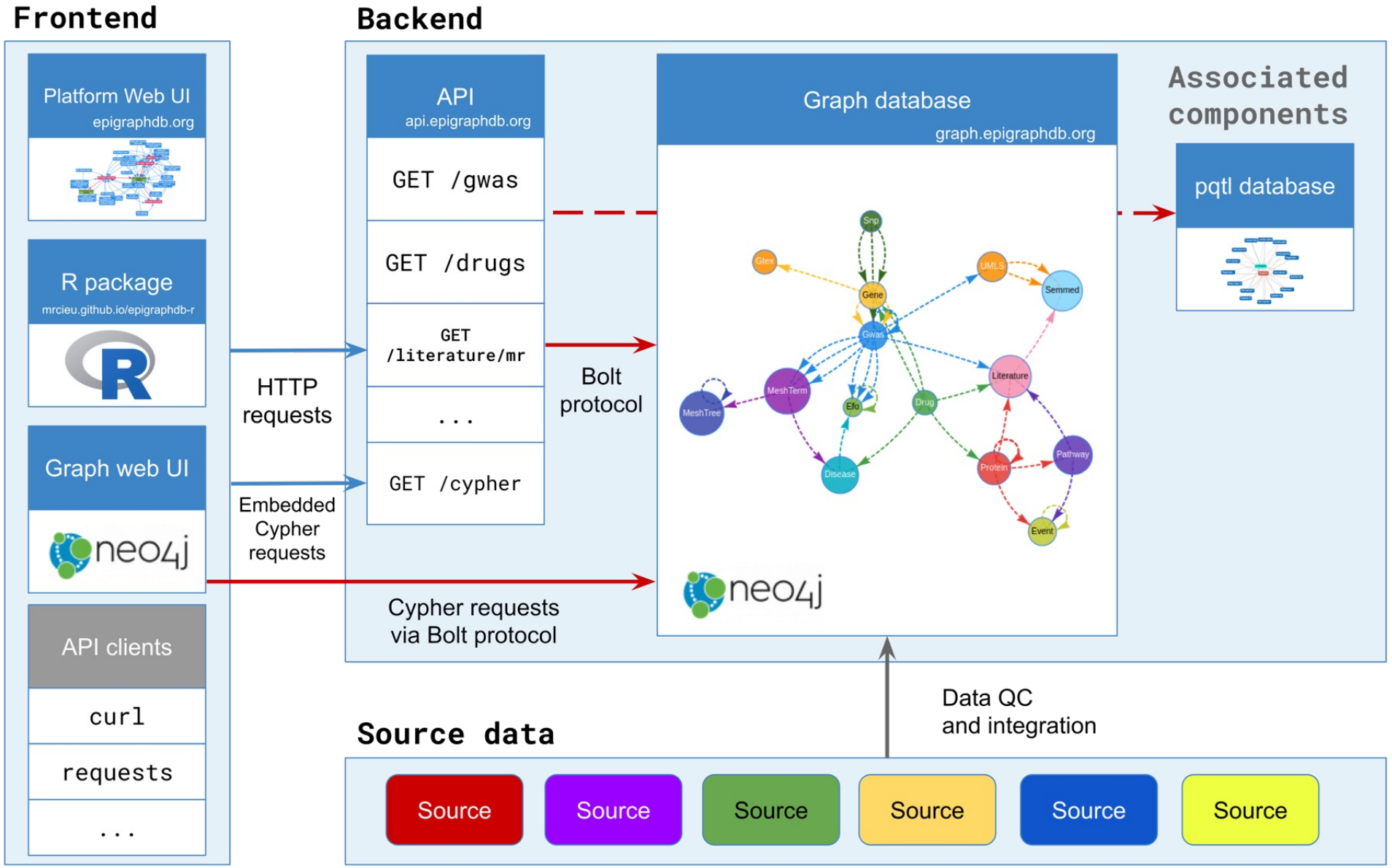
Summary

EpiGraphDB allows the users to search for literature evidence.




Platform

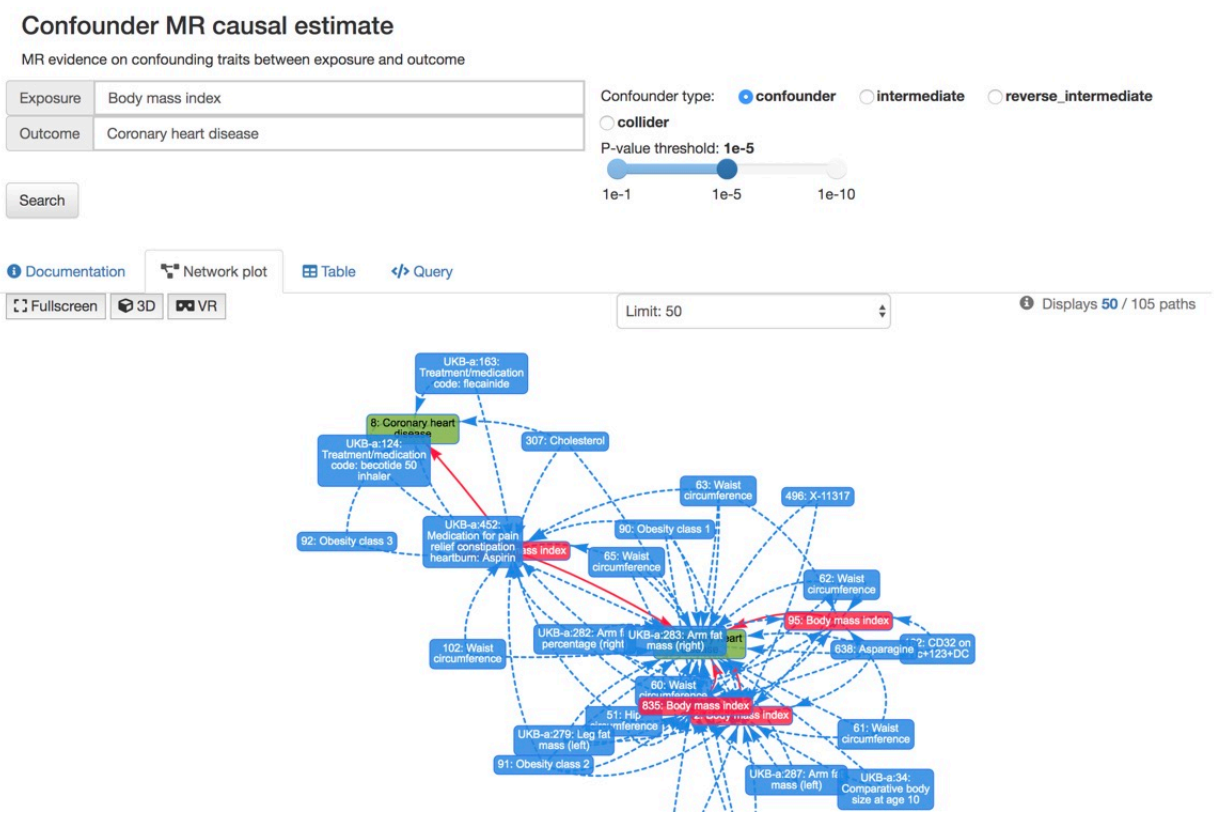
EpiGraphDB platform



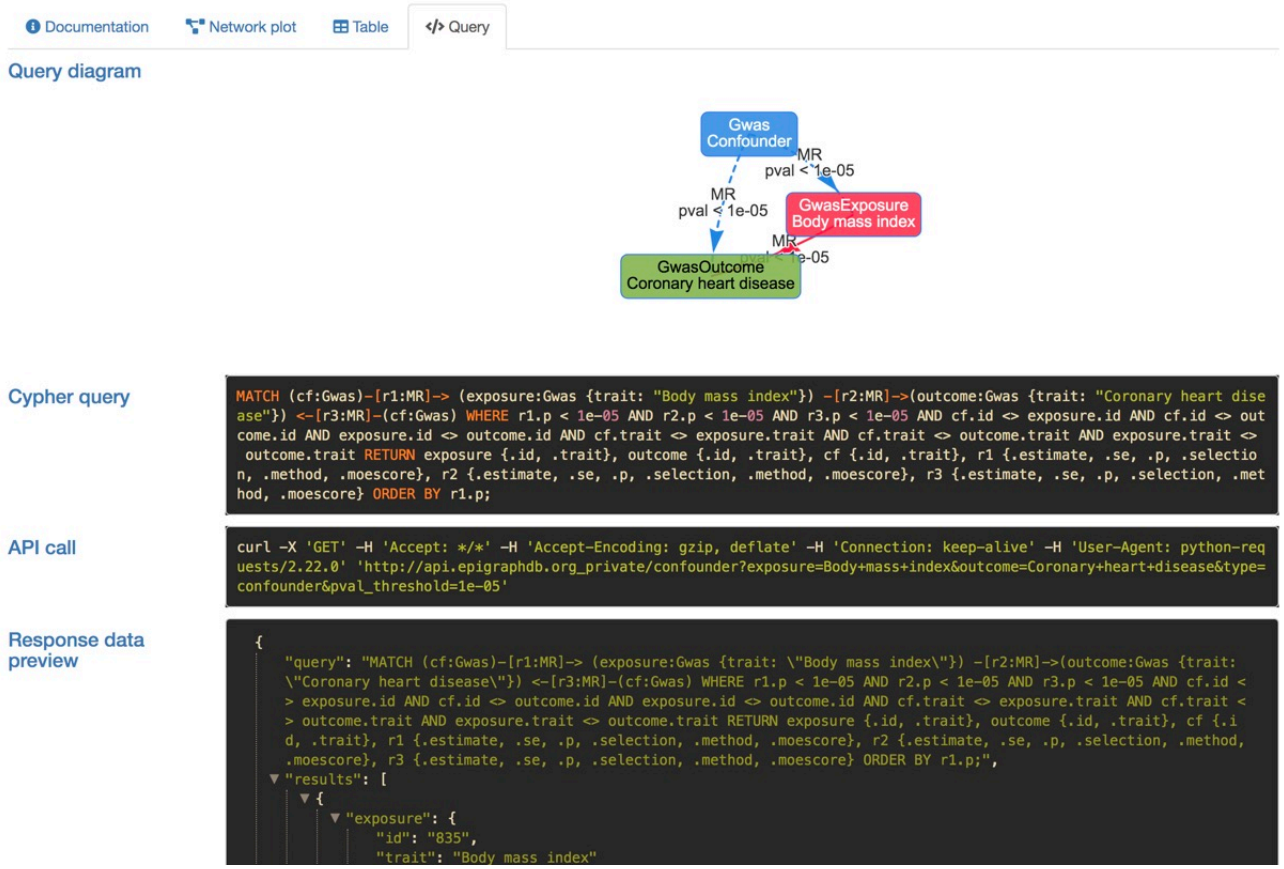
Web UI

http://epigraphdb.org

Network plot



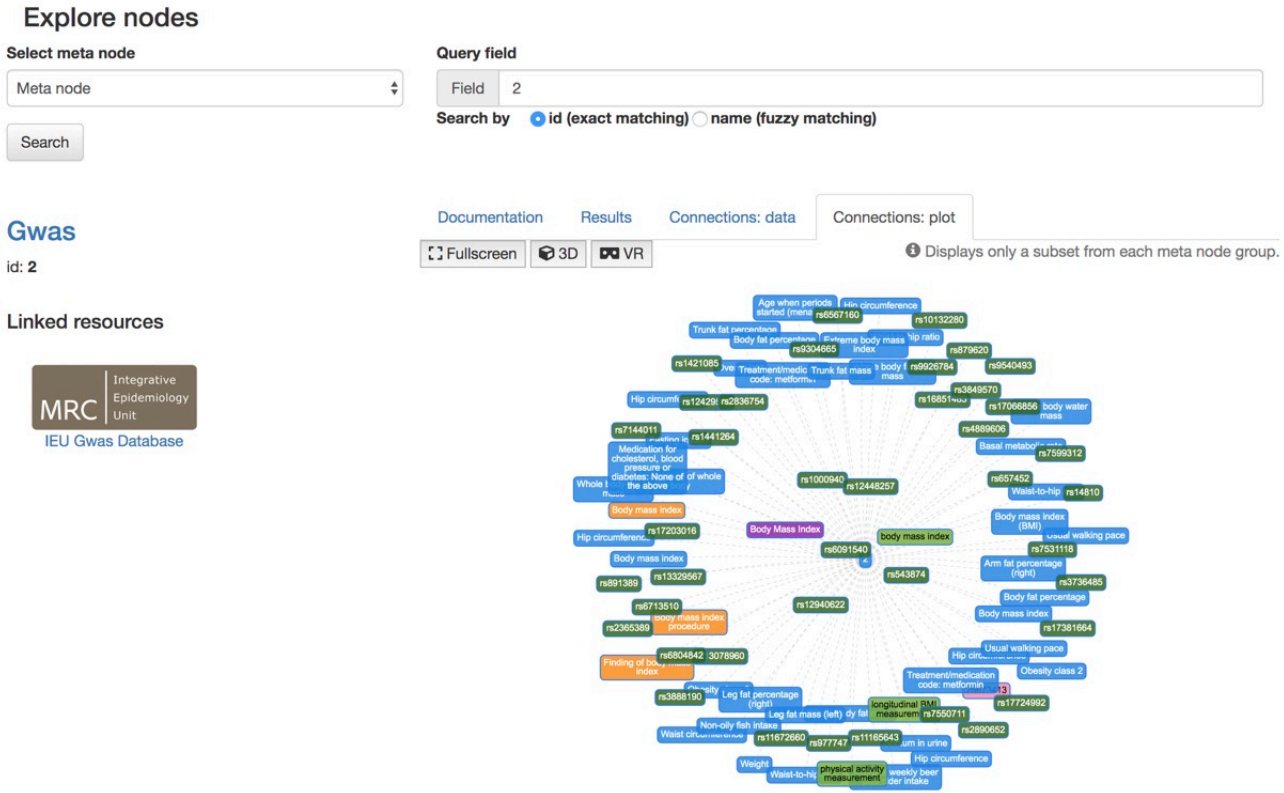
Query



Web UI

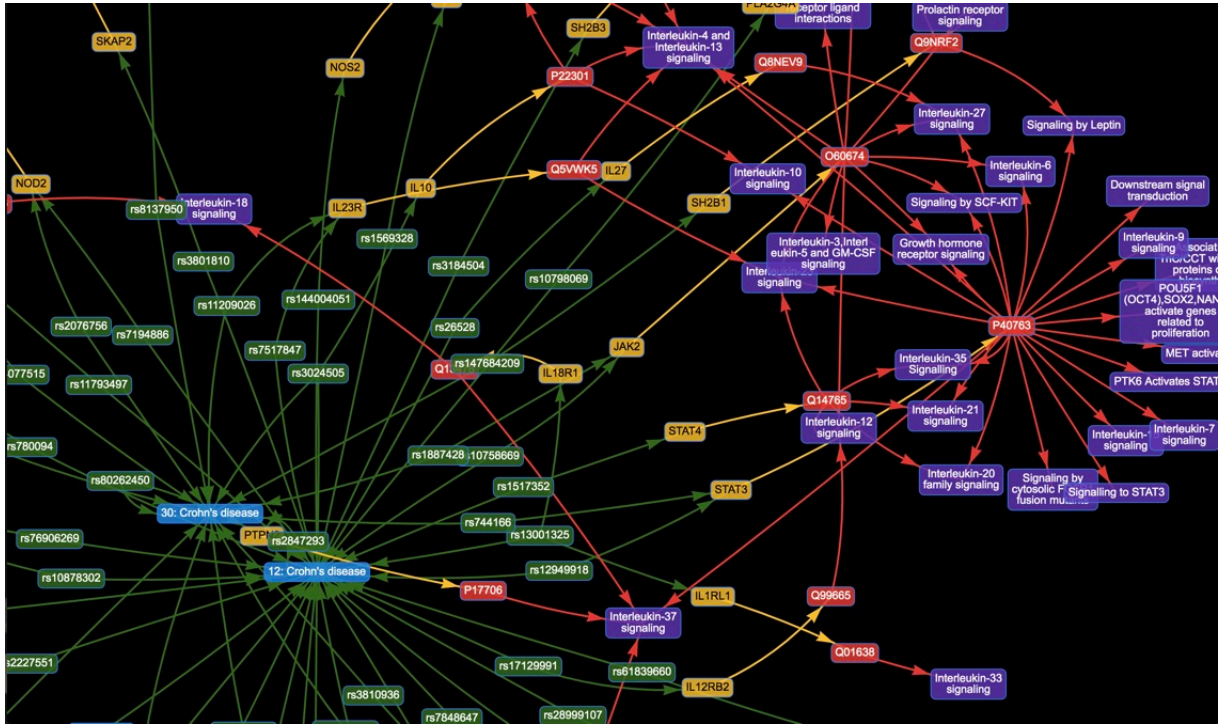
Explorer

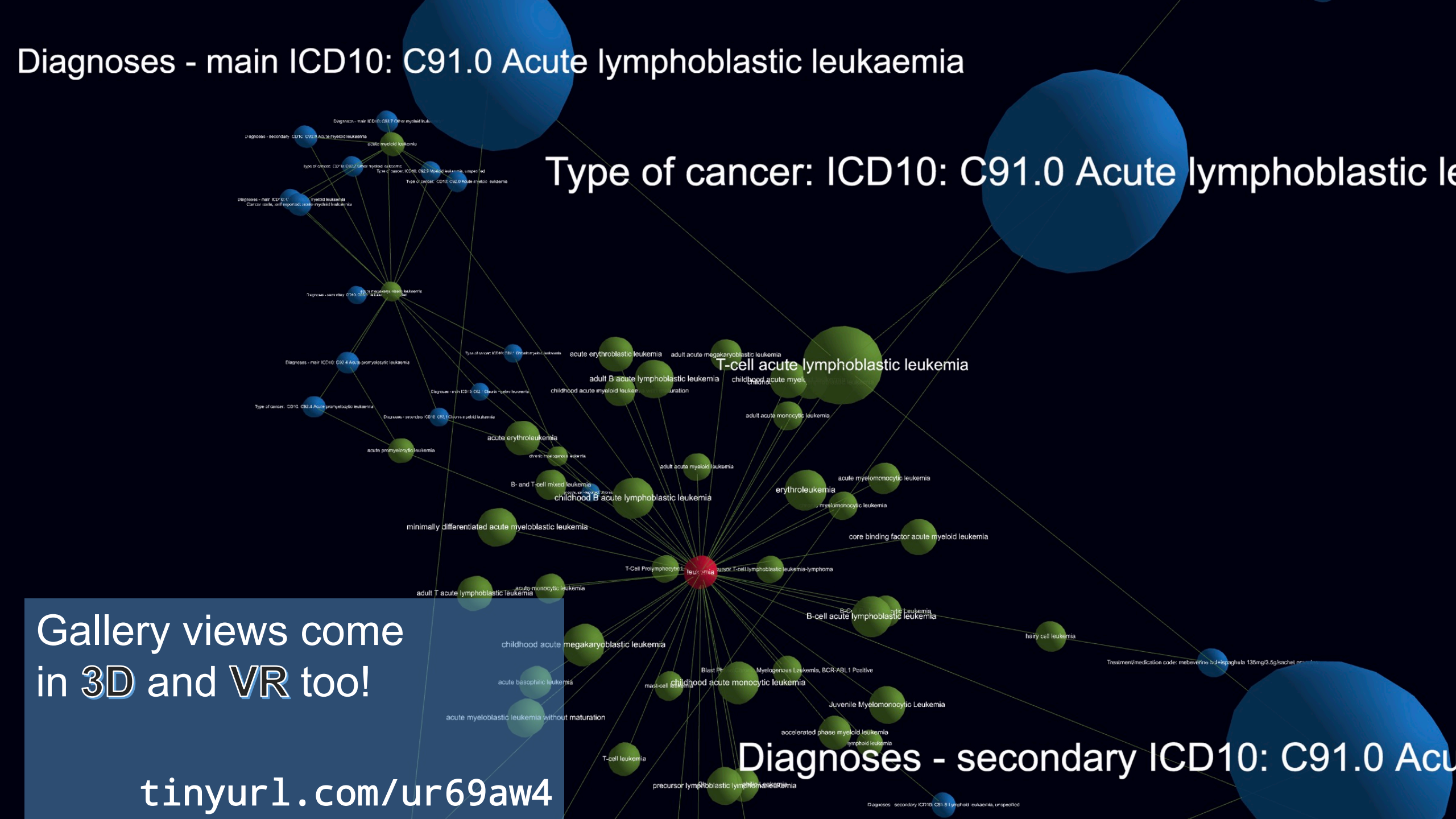
http://epigraphdb.org/explorer



Gallery

http://epigraphdb.org/gallery



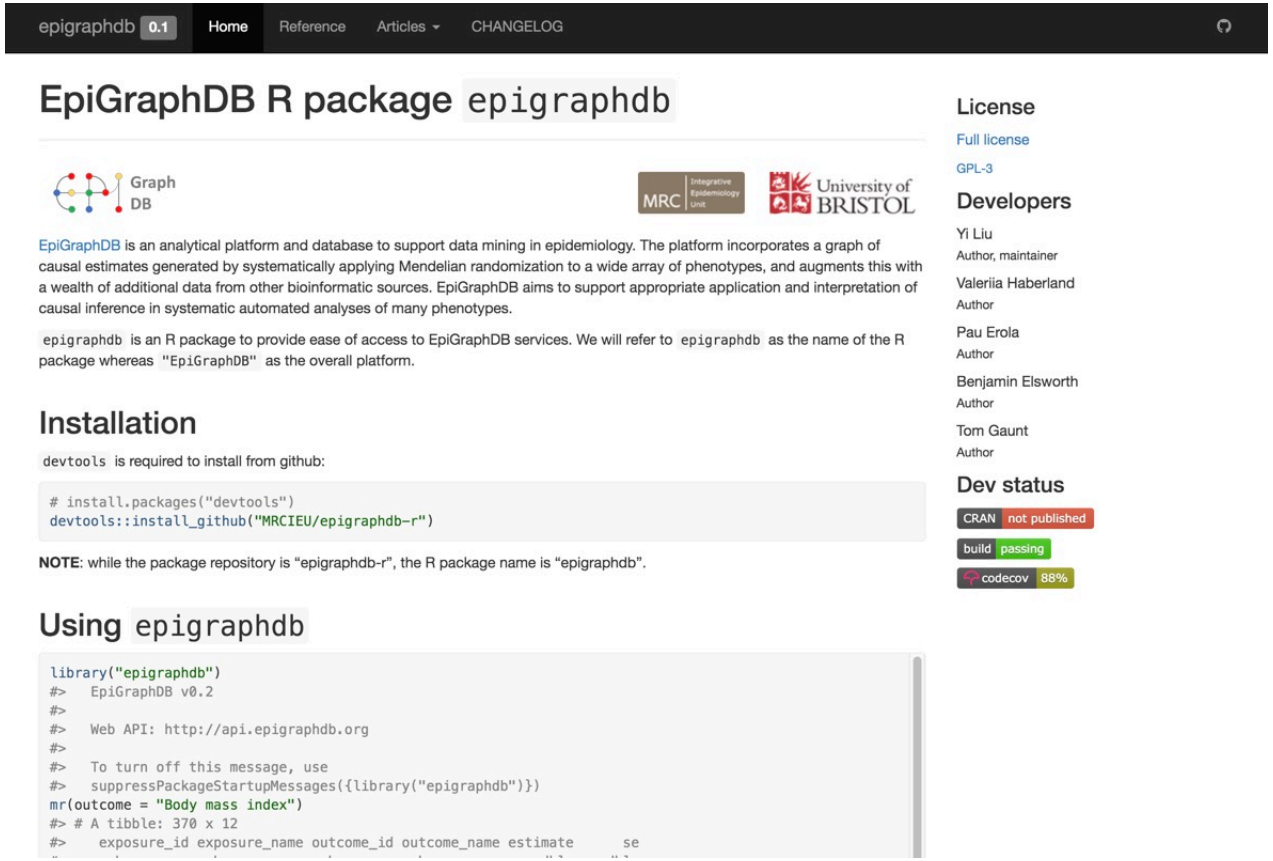

Diagnoses - main ICD10: C91.0 Acute lymphoblastic leukaemia
Type of cancer: ICD10: C91.0 Acute lymphoblastic le
T-cell acute lymphoblastic leukemia
acute erythroblastic leukemia
adult acute megakaryoblastic leukemia
adult B acute lymphoblastic leukemia
childhood acute myeloid leuker
uration
adult acute monocytic leukemia
acute erythroleukemia
acute promyelocytic leukemia
chronic myelogenous leukemia
adult acute myeloid leukemia
B- and T-cell mixed leukemia
childhood B acute lymphoblastic leukemia
erythroleukemia
acute myelomonocytic leukemia
myelomonocytic leukemia
minimally differentiated acute myeloblastic leukemia
core binding factor acute myeloid leukemia
T-Cell Prolymphocytic L
leuk emia
ursor T-cell lymphoblastic leukemia-lymphoma
adult T acute lymphoblastic leukemia
acute monocytic leukemia
B-cell acute lymphoblastic leukemia
hairy cell leukemia
childhood acute megakaryoblastic leukemia
acute basophilic leukemia
acute myeloblastic leukemia without maturation
Blast Ph
Myelogenous Leukemia, BCR-ABL1 Positive
mast-cell leukemia
childhood acute monocytic leukemia
Juvenile Myelomonocytic Leukemia
accelerated phase myeloid leukemia
lymphoid leukemia
T-cell leukemia
precursor lymphoblastic lymphoma/leukemia
Diagnoses - secondary ICD10: C91.0 Acu
Diagnoses - secondary ICD10: C91.0 Lymphoid leukaemia, unspecified
Gallery views come
in 3D and VR too!
tinyurl.com/ur69aw4

epigraphdb R package

https://mrcieu.github.io/epigraphdb-r

- API client package
- `tidyverse` compliant
- `pkgdown` doc site
- Need user feedbacks!

```
install.packages("devtools")
devtools::install_github(
  "MRCIEU/epigraphdb-r"
)
library("epigraphdb")
```



Query EpiGraphDB: colliders

`epigraphdb::confounder` R package

```
confounder(
  exposure = "Body mass index",
  outcome = "Coronary heart disease",
  type = "collider"
)
```

`epigraphdb::query_epigraphdb`

```
query_epigraphdb(
  endpoint = "/confounder",
  params = list(
    exposure = "Body mass index",
    outcome = "Coronary heart disease",
    type = "collider"
  )
)
```

Python `requests` API

```
import requests

api_url = "http://api.epigraphdb.org
endpoint = "/confounder"
response = requests.get(
    url=api_url + endpoint,
    params={
        "exposure": "Body mass index",
        "outcome": "Coronary heart
disease",
        "type": "collider"
    }
)

print(response.json())
```

Documentation (pkgdown):
https://mrcieu.github.io/epigraphdb-r

Documentation (swagger interface):
http://api.epigraphdb.org

Summary

Summary

http://docs.epigraphdb.org/slides/2019-12-ieu-meeting.pdf

- EpiGraphDB as a database and platform powers the the data mining of high dimensional, complex epidemiological relationships.
- We are actively developing EpiGraphDB and working on associated research studies.
- Please use EpiGraphDB and get in touch!
- Email: feedback@epigraphdb.org

Acknowledgements

EpiGraphDB

Yi Liu
Benjamin Elsworth
Valeriia Haberland
Pau Erola
Jie Zheng
Matt Lyon
Tom R Gaunt

pQTL project

Jie Zheng
Valeriia Haberland
Benjamin Elsworth
Denis Baird
Venexia Walker
Tom Richardson
Kurt Taylor
James Staley
George Davey Smith
Philip Haycock
Gibran Hemani
Robert Scott
Biogen & GSK
collaborators

Reference

- Elsworth, B., et al., 2018. MELODI: mining enriched literature objects to derive intermediates.
- Elsworth, B., et al., forthcoming a. IEU GWAS database https://gwas.mrcieu.ac.uk/
- Elsworth, B., et al., forthcoming b. Vectology: exploring biomedical variable relationships using sentence embedding and vectors.
- Hemani, G., et al., 2017. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *BioRxiv*, p.173682.
- Richardson, T.G., et al., 2019. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife*, *8*, p.e43657.
- Zheng, J. et al., 2019. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *BioRxiv*, doi: https://doi.org/10.1101/627398
- Zheng, J., et al., forthcoming. Systematic Mendelian randomization and colocalization analyses of the plasma proteome and blood transcriptome to prioritize drug targets for complex disease.