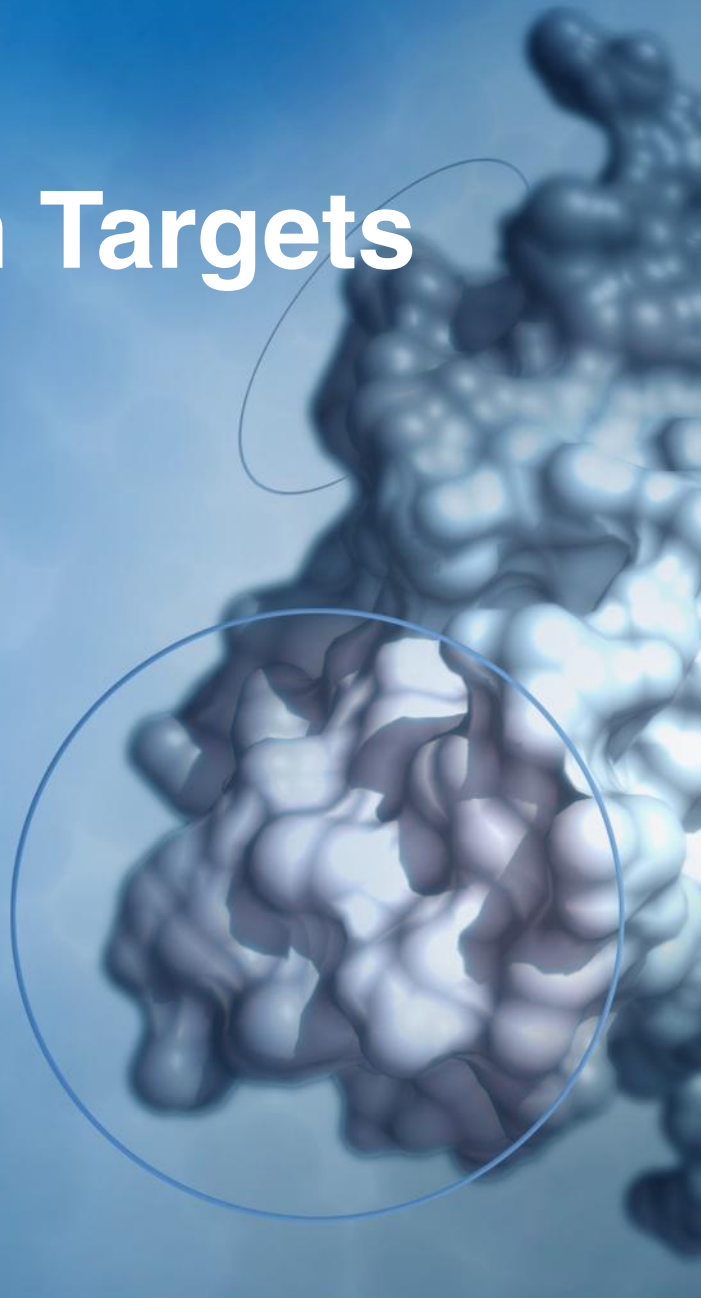
Mining Gene-Disease Associations with Open Targets

Alzheimer's Research UK
Cambridge DDU

Denise Carvalho-Silva

Wellcome Genome Campus, United Kingdom
Open Targets Consortium
Core Bioinformatics team

Open Targets



Materials

https://github.com/deniseOme/training

slides

coursebook

Course's objectives

What is the **Open Targets Consortium**?

What is the **Open Targets Platform**?

How to **navigate** the Platform?

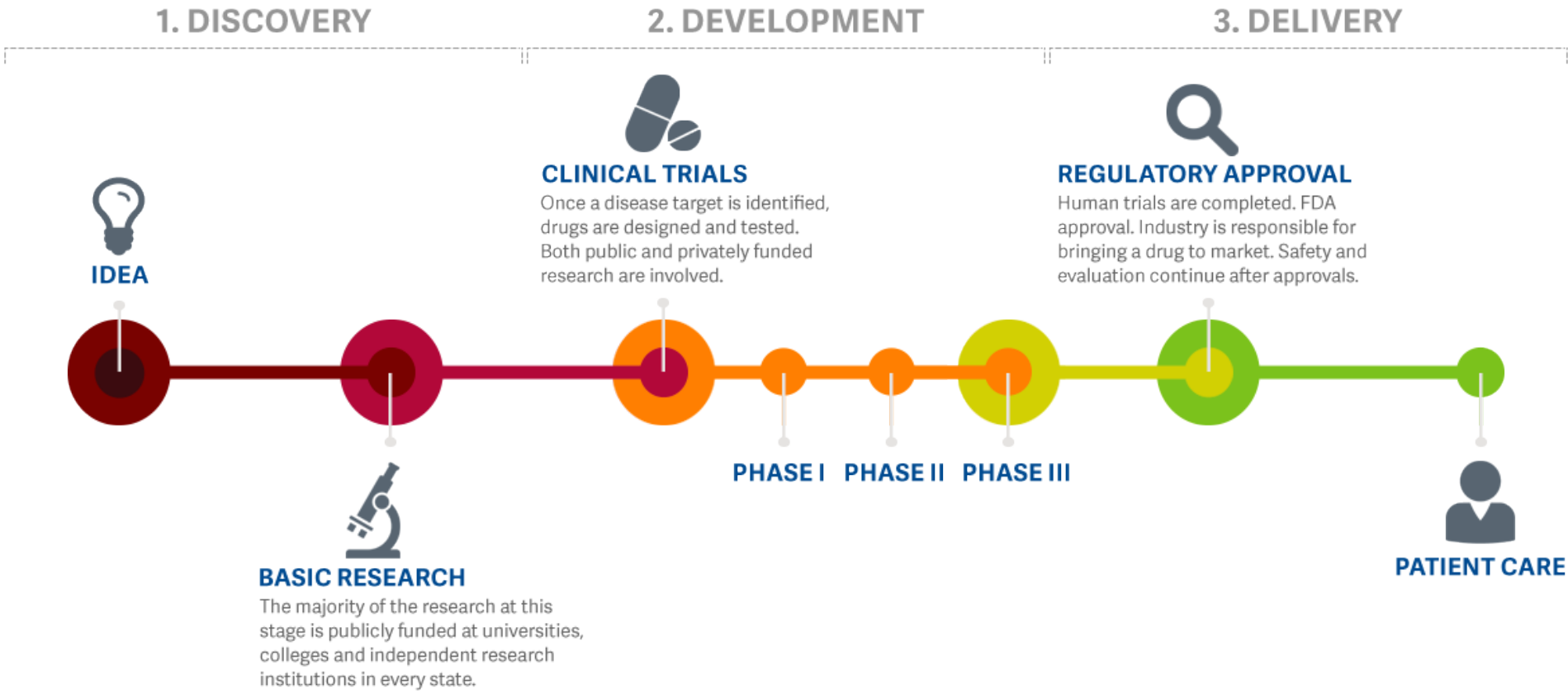
How to **connect** with the team

Outline

- The Open Targets Consortium
- The Open Targets Platform
- Live demos and hands on
- Get in touch

Open Targets

Drug discovery path: timeline



Lengthy, costly, low success rate, **HIGH ATTRITION RATES**

Professor Sir
Mike Stratton
Director, Sanger Institute

Can we improve target identification?

Yes, we can!
And we should.

Patrick Vallance, President
Pharmaceuticals R&D
GlaxoSmithKline

Professor Dame
Janet Thornton
former Director, EMBL-EBI

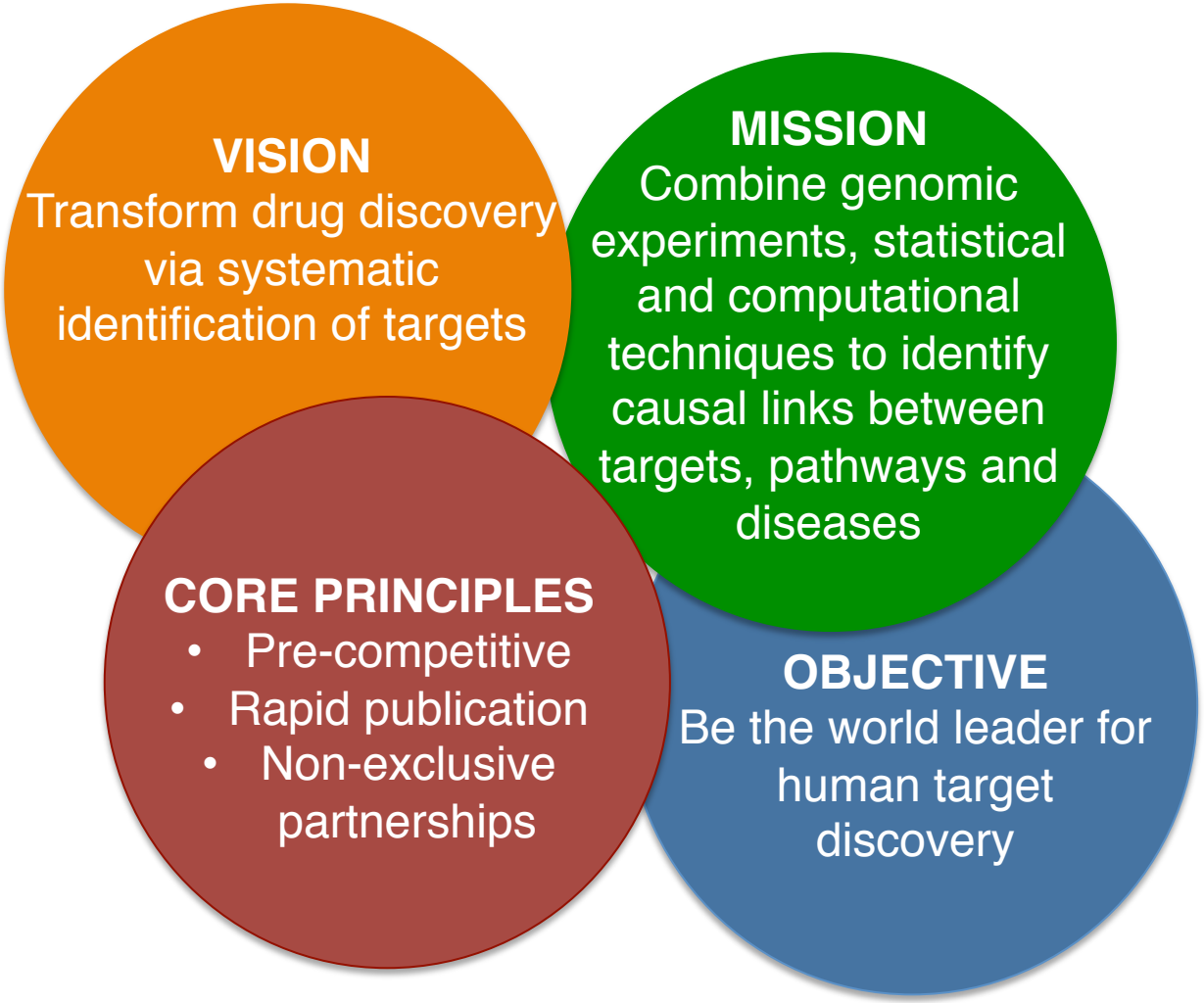
But one institution can not do it alone.

gsk

EMBL-EBI

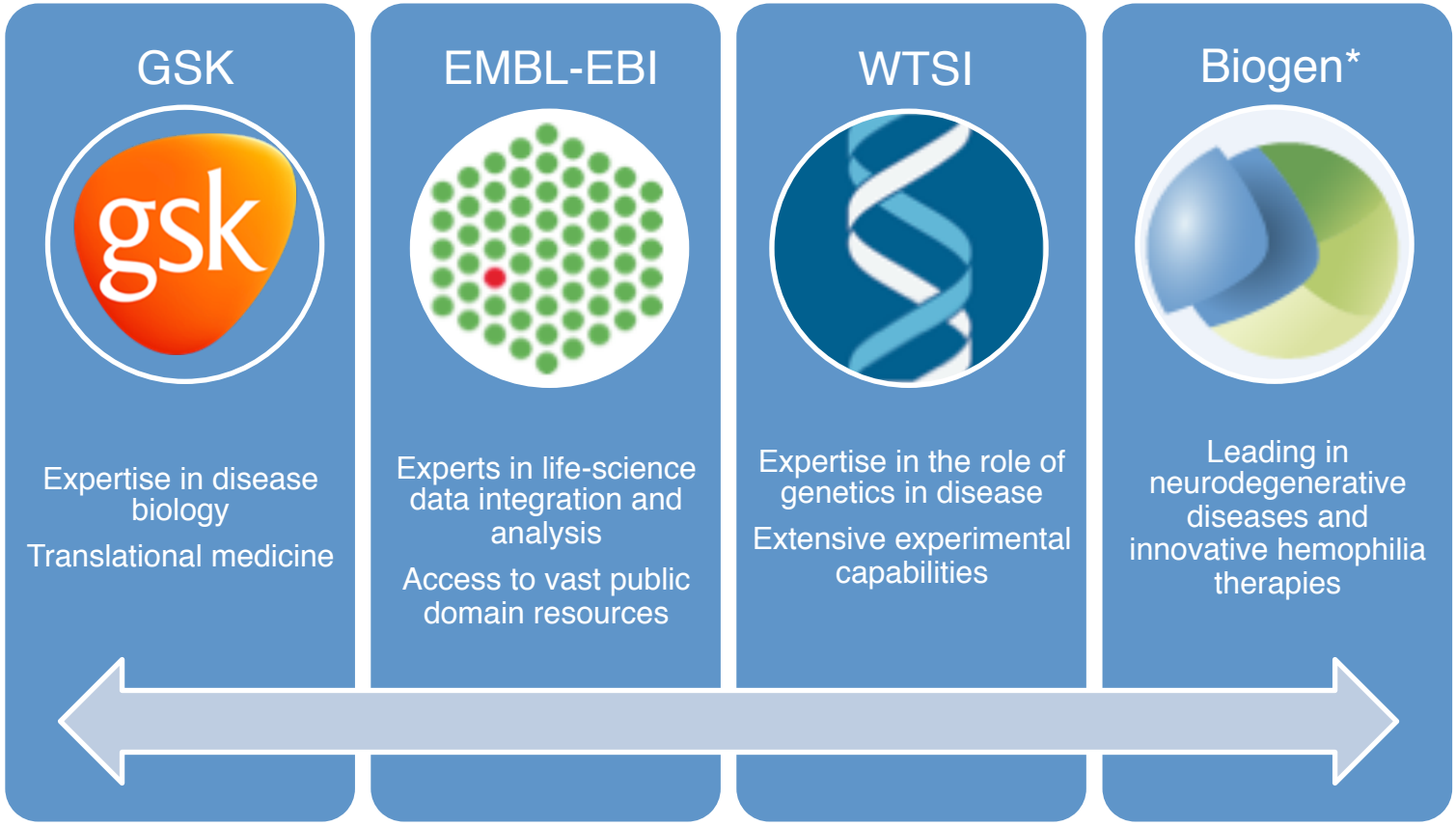
wellcome trust sanger institute

Open Targets Consortium*



* Three founding partners, March 2014

Who is Open Targets today?



* Biogen joined the consortium in February 2016

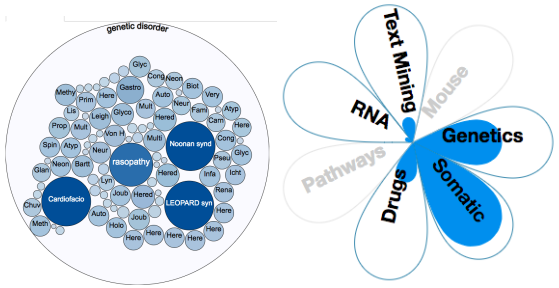
Open Targets

Two major areas of work in Open Targets

Experimental projects

Generate new evidence
CRISPR/Cas9, Organoids
(cellular disease models)

Core bioinformatics pipelines



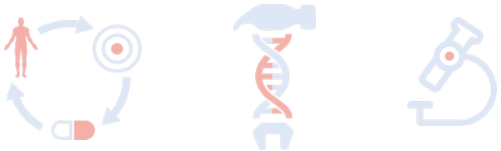
Integration of available data
Web portal
REST API and data dumps

Concurrent
www.opentargets.org/projects

Open Targets

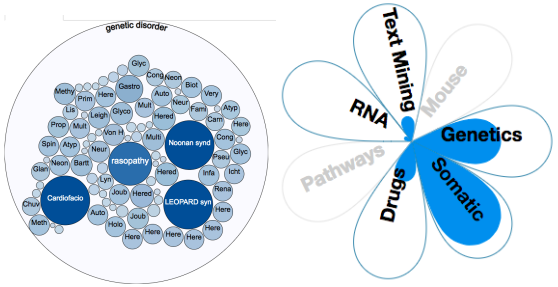
Two major areas of work in Open Targets

Experimental projects



Generate new evidence
CRISPR/Cas9, Organoids
(cellular disease models)

Core bioinformatics pipelines



Database for data integration
Web portal
REST API and data dumps

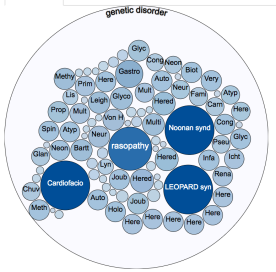
Concurrent
www.opentargets.org/projects

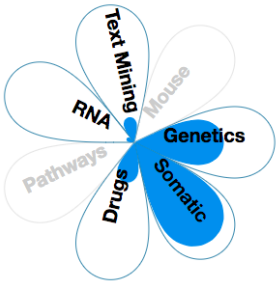
Open Targets

Open Targets Platform*

- Developed by the Core Bioinformatics team at EMBL-EBI
- Allow users to identify target–disease associations
- Improvements driven by you

https://www.targetvalidation.org/



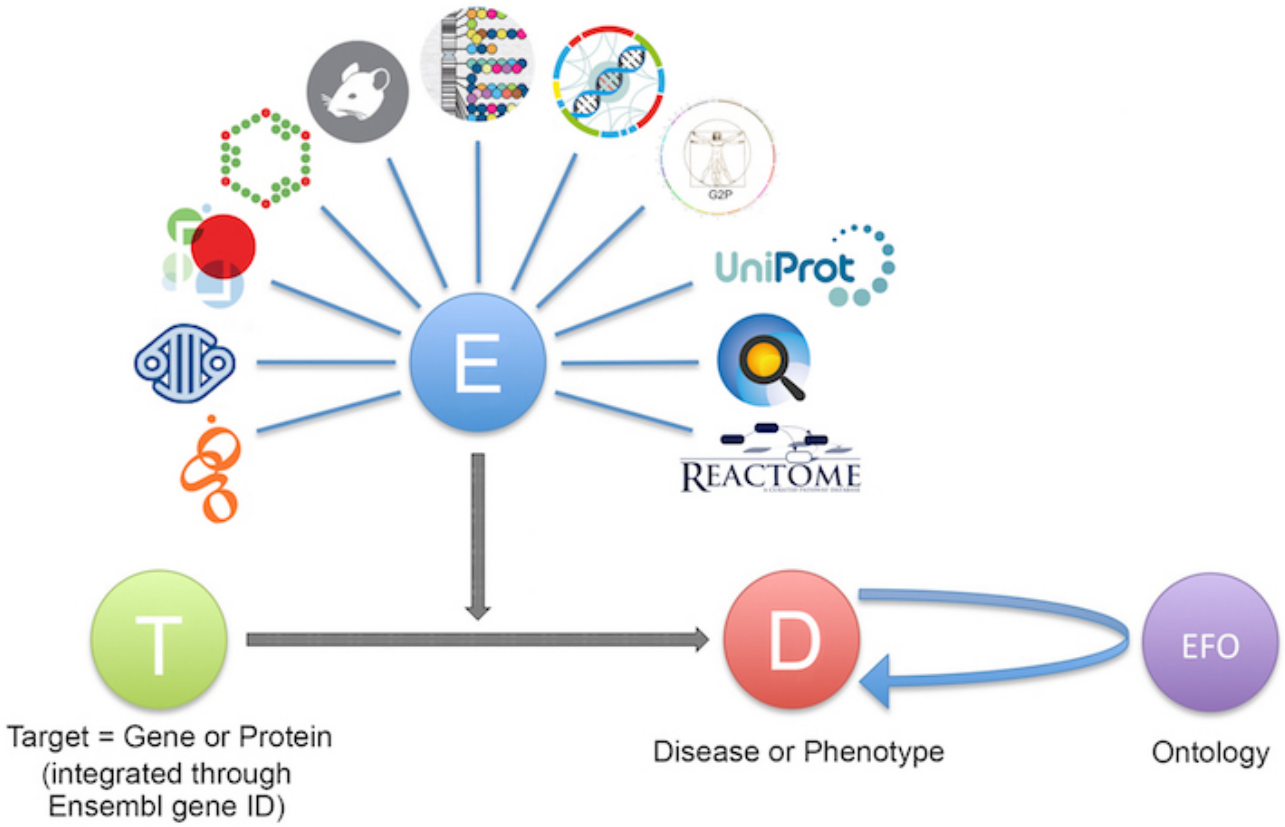


Open Targets

* First release: December 2015

Open Targets Platform

Evidence model for target–disease associations

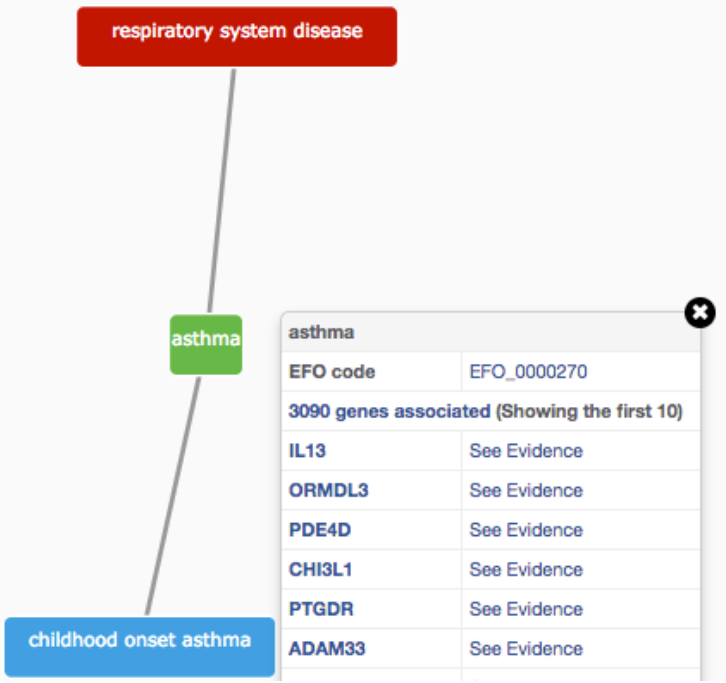


https://github.com/opentargets/json_schema/blob/master/doc/instructions.md

Open Targets

Experimental Factor Ontology* (EFO)

- Ontology: smart dictionary → relationships between entities
- EFO: way to organise experimental variables (e.g. diseases)



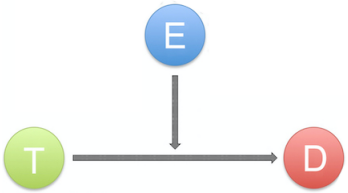
controlled vocabulary
+
hierarchy (relationship)

* https://www.ebi.ac.uk/efo/

Increases the richness of annotation
Promotes consistency
Allow for easier and automatic integration

Evidence from publicly available data

- Similar data sources are grouped into data types

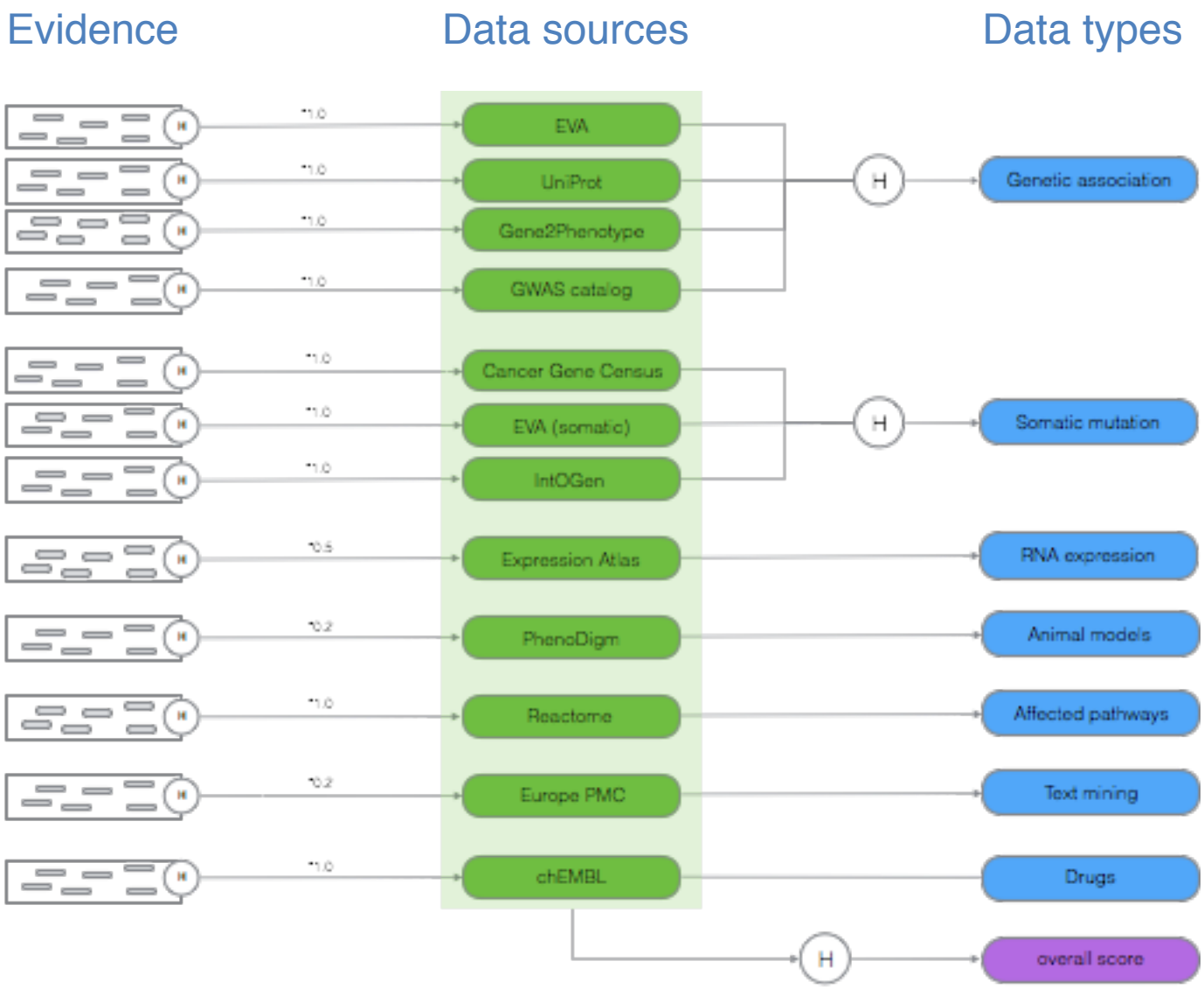


Data sources	Data types
GWAS catalog, UniProt, EVA, G2P	Genetic associations
Cancer Gene Census, EVA, IntOgen	Somatic mutations
Expression Atlas	RNA expression
ChEMBL	Drugs
Reactome	Affected pathways
Europe PMC	Text mining
PhenoDigm	Animal models
Your favourite data?	**Let us know!**

How do we associate diseases and phenotypes w/ targets?

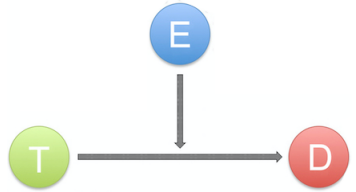
1. ChEMBL, UniProt, EVA (w/ ClinVar) curate diseases and phenotypes
2. Map disease/phenotypes to an ontology using EFO and HPO terms
3. Use genes as proxies for our targets
4. Create target-disease evidence JSON objects
5. Calculate for each supporting evidence the likelihood of gene A being associated with disease B
6. Compute integrated target-disease scores at the levels of data source, data type and overall score

Score approach and aggregation

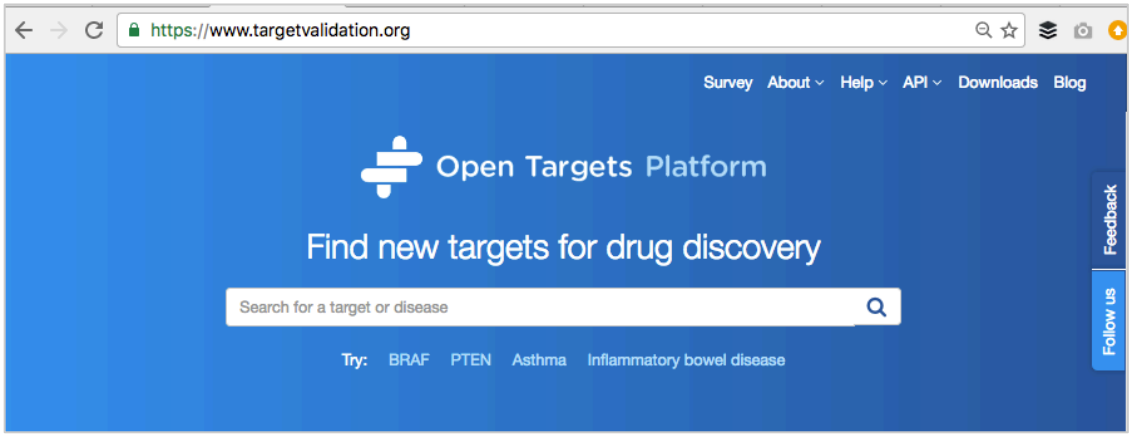


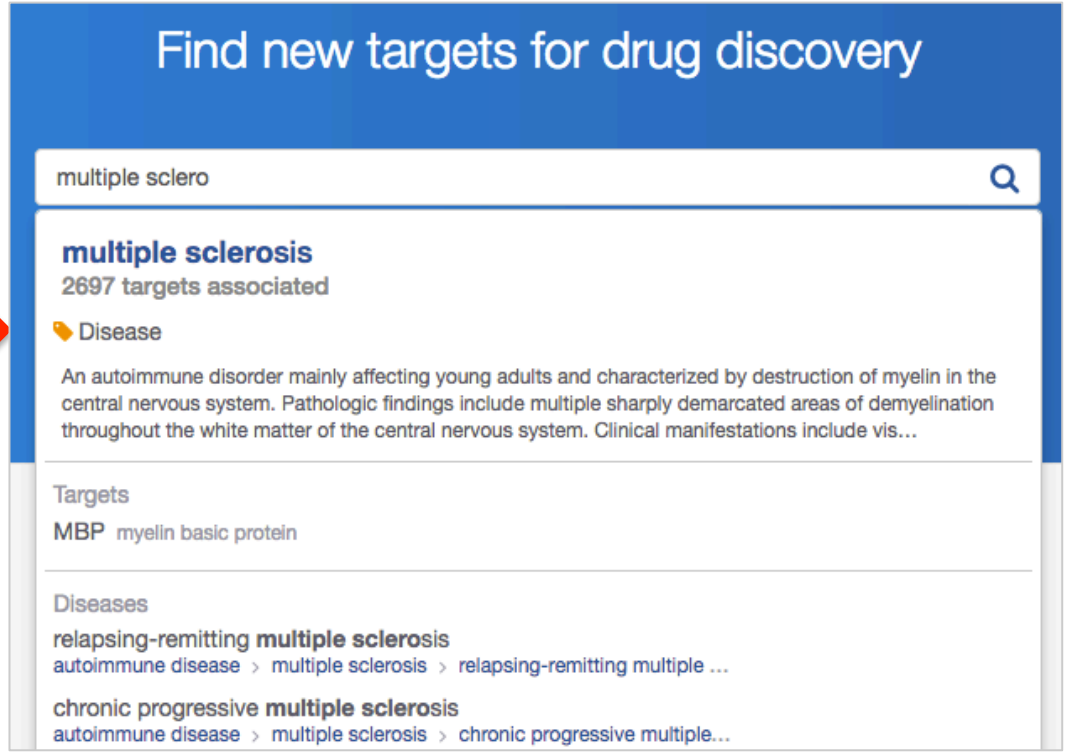
Evidence score = $f * S * C$ (frequency x Severity x Confidence)

$H = S_1 + S_2/2^2 + S_3/3^2 + S_4/4^2 + S_i/i^2$



Which targets are associated with a disease?





Demo 1:

screenshots: next slides
coursebook: pages 9-15

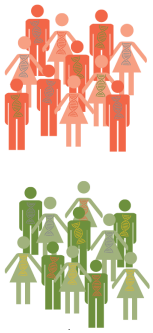
https://www.targetvalidation.org/

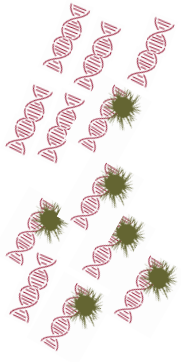
Open Targets

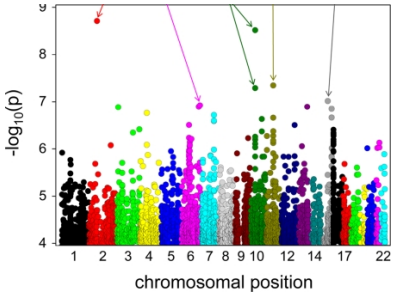

Open Targets

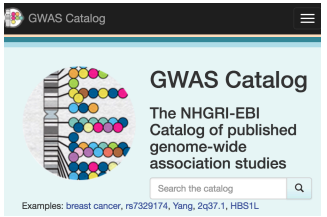
Data sources: GWAS catalog

- Genome Wide Association Studies
- Array-based chips → genotyping 100,000 SNPs genomewide









Open Targets

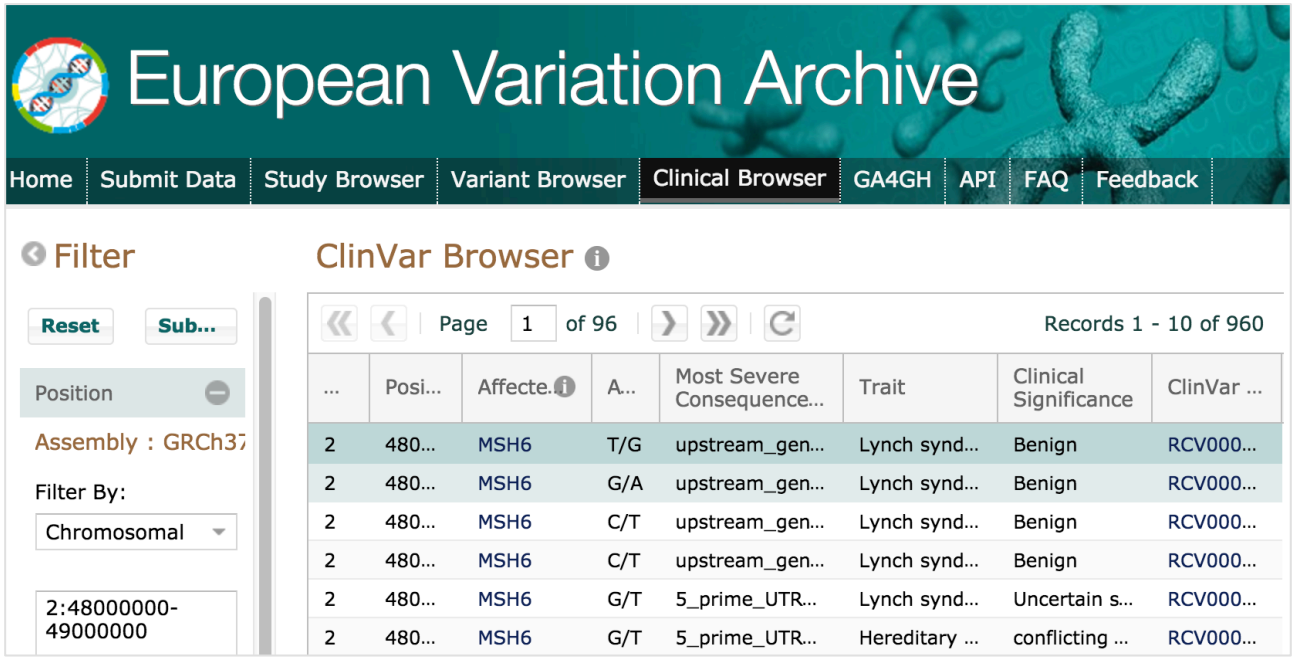
Data sources: UniProt

- Protein: sequence, annotation, function

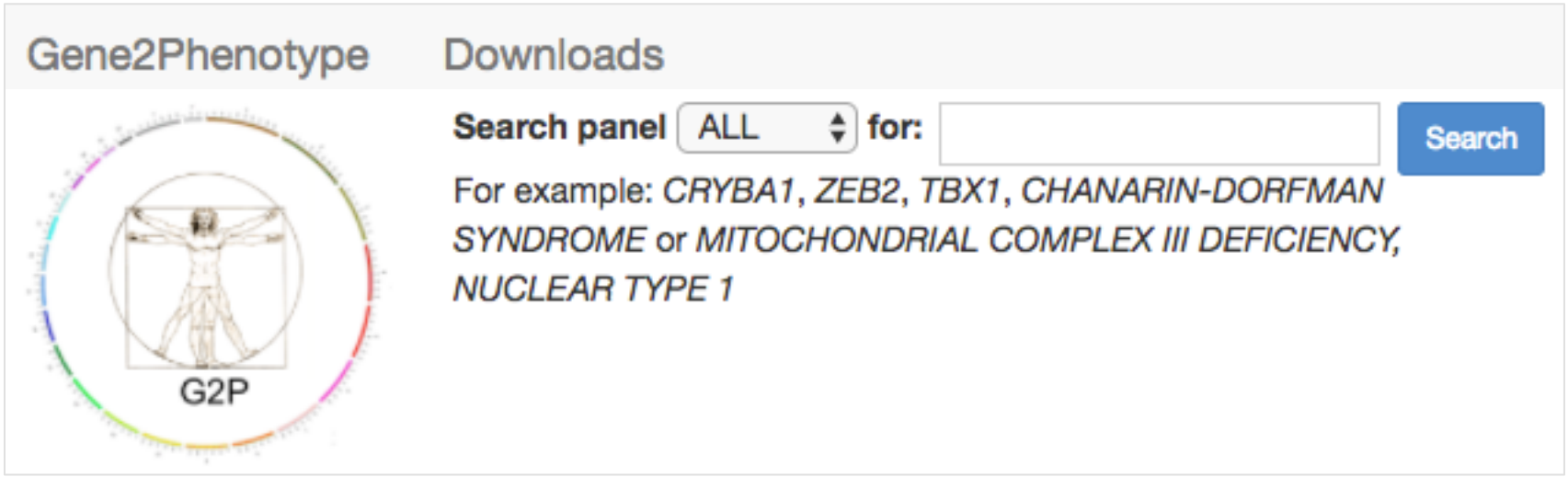
- Manual curation of coding variants, seen in patients

Data sources: EVA

- Germline and somatic variants
- With ClinVar information for rare diseases



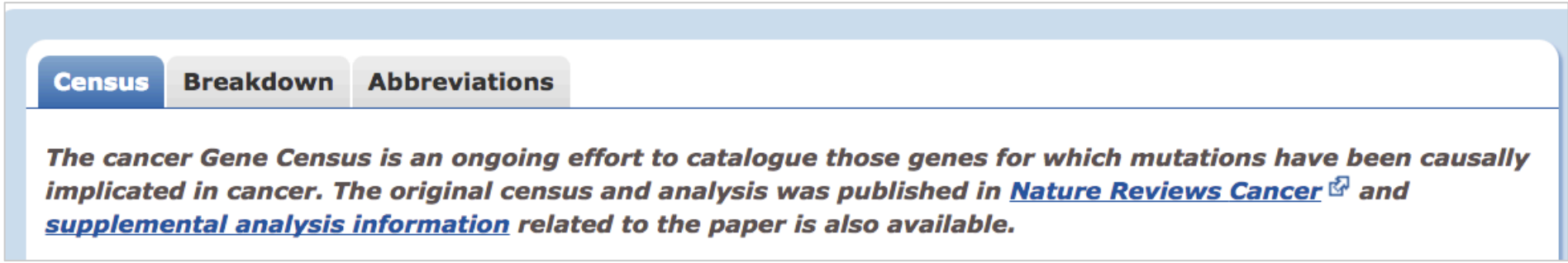
Data sources: Gene2Phenotype



- Variants, genes, phenotypes in rare diseases
- Literature curation → consultant clinical geneticists in the UK

Open Targets

Data sources: The Cancer Gene Census



- Genes with mutations causally implicated in cancer
- Gene associated with a cancer plus other cancers associated with that gene

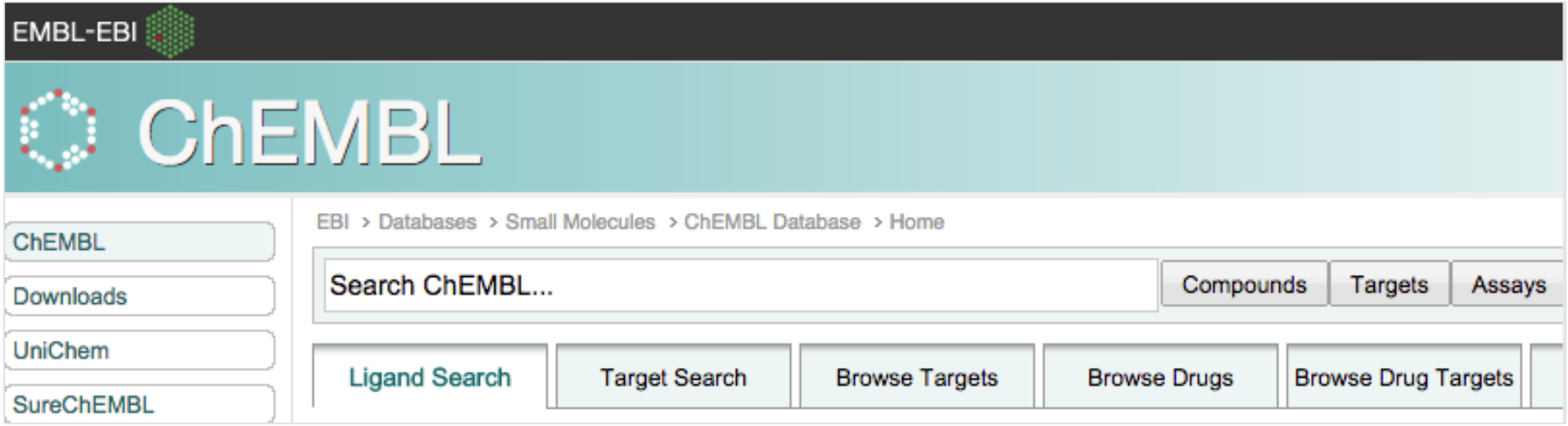
Open Targets

Data sources: IntOGen

- Genes and somatic (driver) mutations
- Involvement in cancer biology

Open Targets

Data sources: ChEMBL

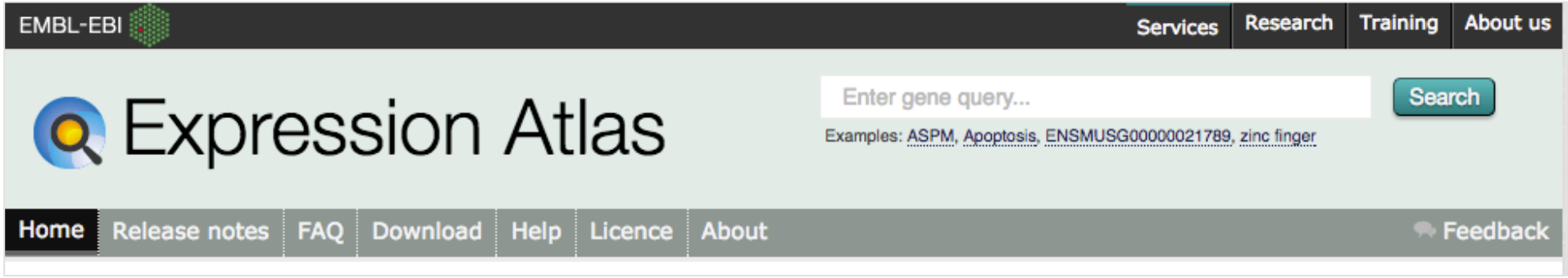


- Known drugs linked to a disease and a known target
- FDA approved for clinical trials or marketing

Data sources: Reactome

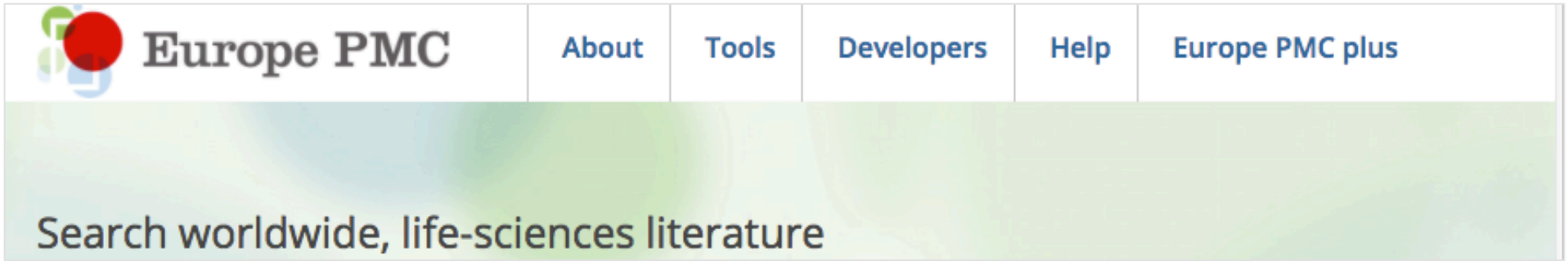
- Biochemical reactions and pathways
- Manual curation of pathways affected by mutations

Data sources: Expression Atlas



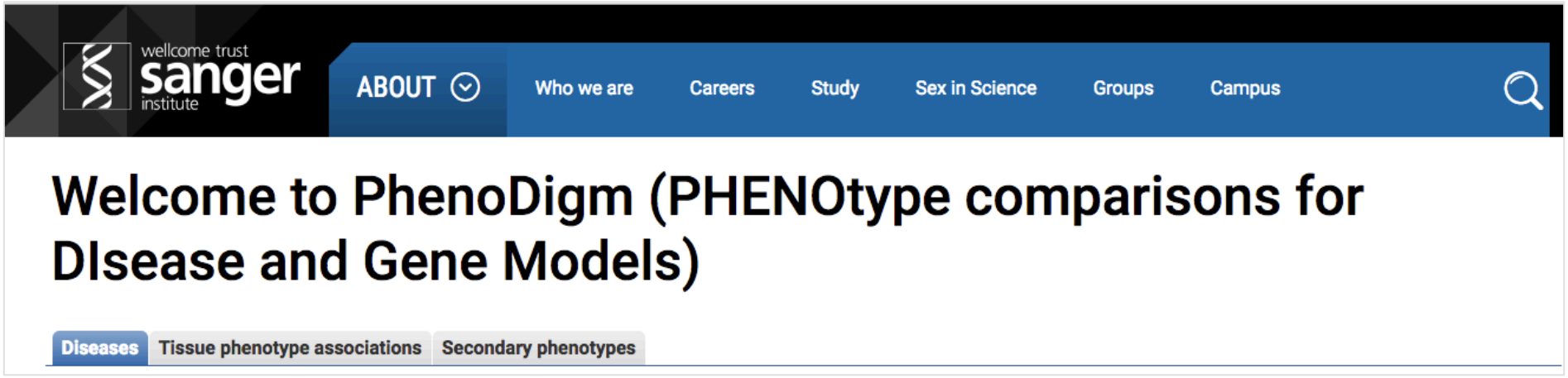
- Baseline expression for human genes
- Differential mRNA expression (healthy *versus* diseased)

Data sources: Europe PMC



- Mining titles, abstracts, full text in research articles
- Target and disease co-occurrence in the same sentence

Data sources: PhenoDigm



- Semantic approach to associate mouse models with diseases

Open Targets

Demo 2: Evidence supporting the *CD86* - multiple sclerosis association

- Which genetic evidence supports this association?
- Can you view this in a genome browser display?
- Are there any drugs in clinical trials for this disease?
- Which cell/tissue has the highest RNA expression(Illumina Body Map data)?
- Are there other diseases of the nervous system associated with this target? Can you export the table with this information? How strong is this association?

*coursebook: pages 16-25

Hands-on exercises

Pages 26-29

How confident can you be of the target-disease associations in Open Targets?

Statistical integration, aggregation and scoring*

A) per evidence (e.g. lead SNP from a GWAS paper)

B) per data source (e.g. GWAS catalog)

C) per data type (e.g. Genetic associations)

D) overall

*https://github.com/opentargets/association_score_methods

Factors affecting the relative strength of an evidence

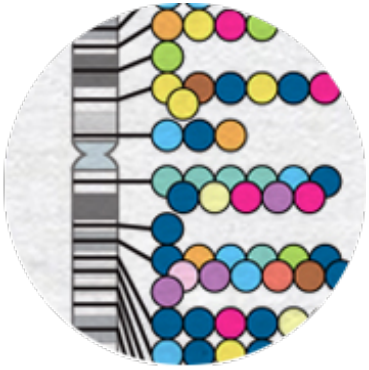
e.g. GWAS Catalog

$$S = f * s * c$$

f, relative occurrence of a target-disease evidence

s, strength of the effect described by the evidence

c, confidence of the observation for the target-disease evidence



f= sample size (cases versus controls)

s = predicted functional consequence

c = *p*-value reported in the paper

Aggregating scores across the data

- Using a mathematical function, the harmonic sum*

$$S_{1..i} = S_1 + \frac{S_2}{2^2} + \frac{S_3}{3^2} + \frac{S_4}{4^2} \ldots + \frac{S_i}{i^2}$$

where S1,S2,...,Si are the individual sorted evidence scores in descending order

- Advantages:
  A) account for replication
  B) deflate the effect of large amounts of data e.g. text mining

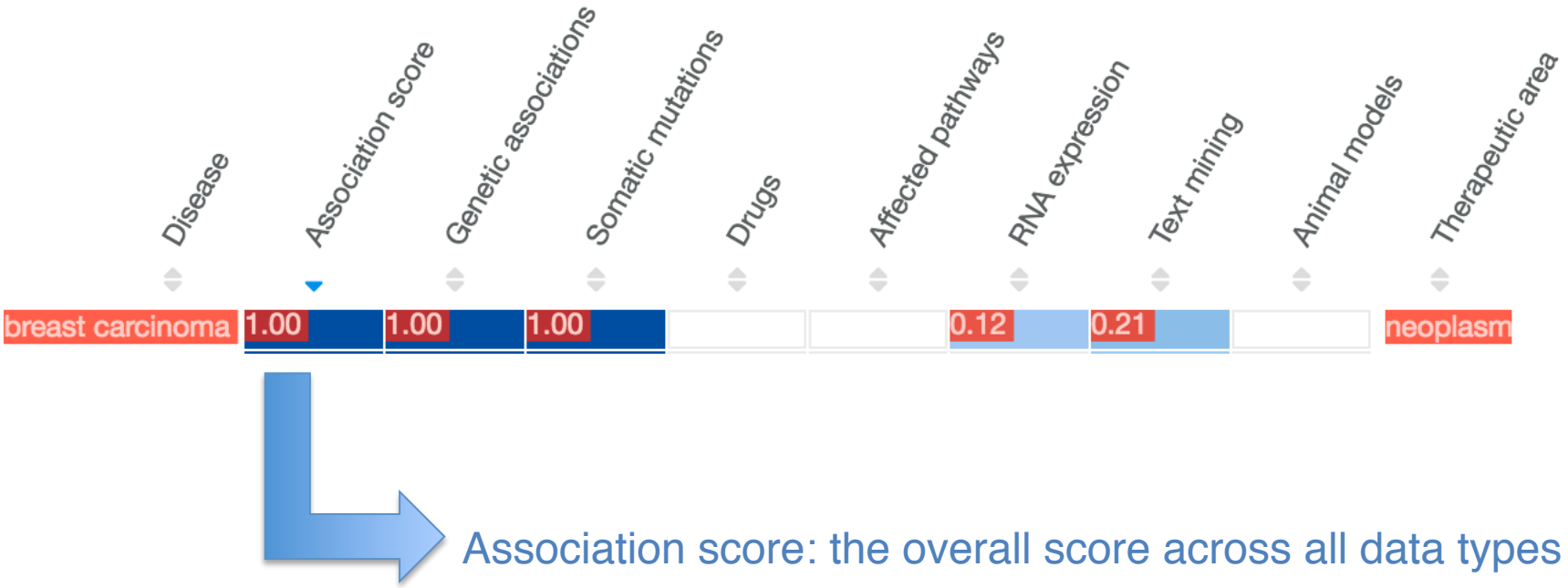
* PMID: 19107201, PMID: 20118918

Disclaimer: score, dos and don’ts

- It’s a ranking of target-disease associations
- It shows how confident we are in the association
- It’s based on data sources, publicly available

- It can help you to design your null hypothesis
- It can help you to decide which target to pursue
- It is NOT sufficient on its own (use it in combination with…)

Ranking the target-disease association



- Based on the data sources
- Different weight applied:

genetic association = drugs = mutations = pathways > RNA expression > animal models = text mining

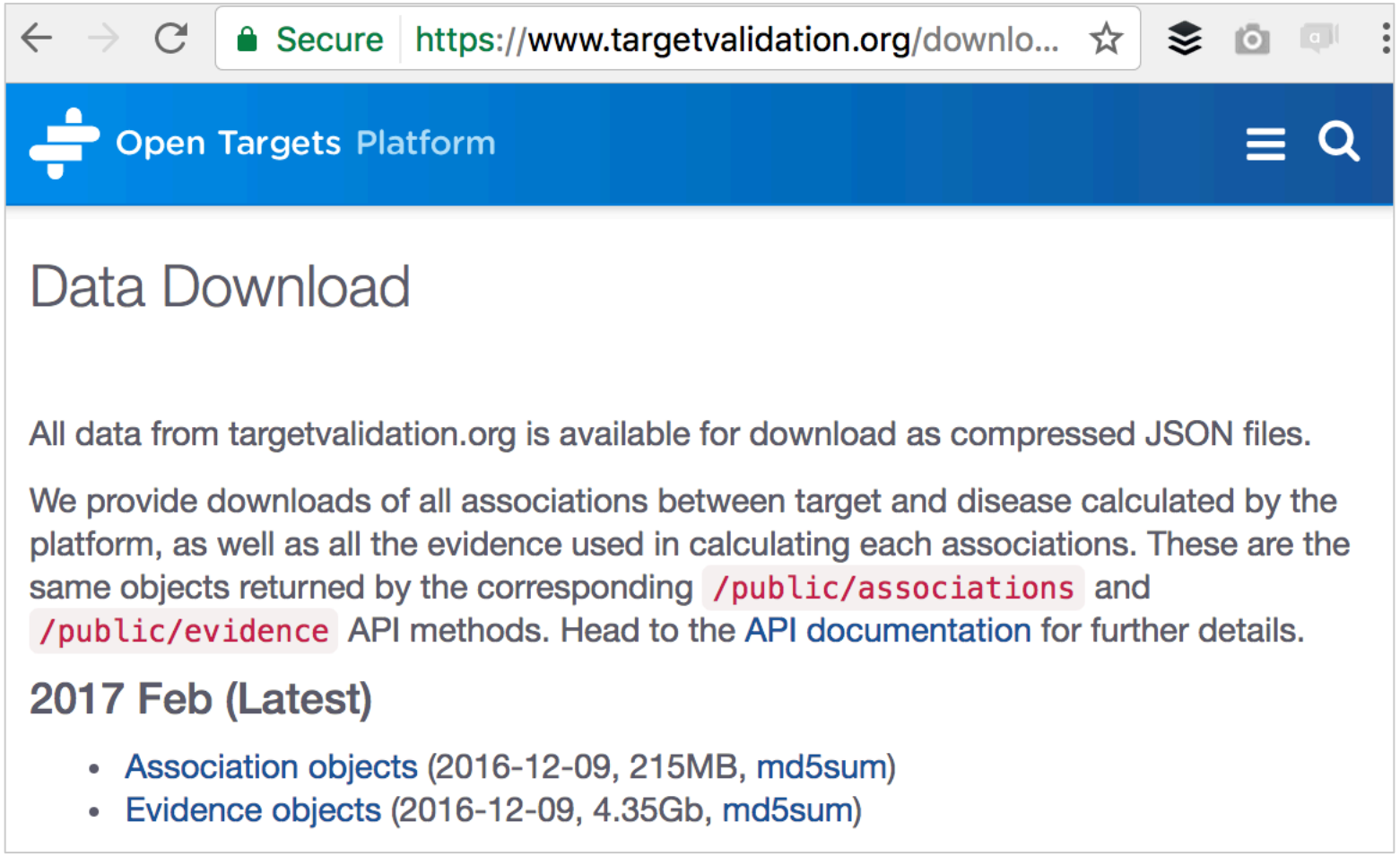
Demo 3: your list of targets for a disease

These are some of the genes associated with Alzheimer's disease from the medical literature: *HFE*, *PSEN1*, *TF*, *APOE*, *ADRB2* , *PSEN2*, and *A2M*.

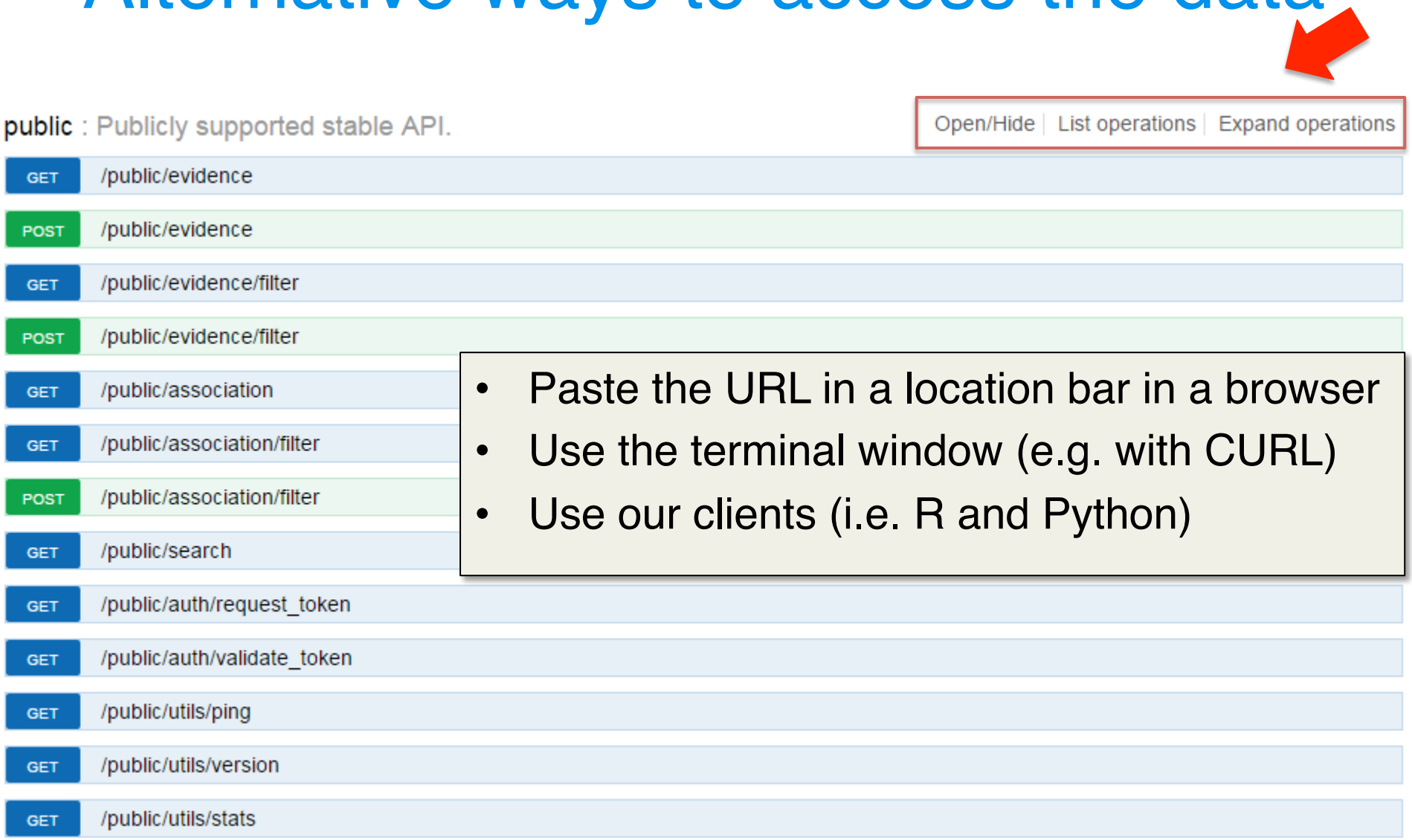
- Which of these have the strongest association w/ Alzheimer's?
- Is there any membrane receptors in this list?
- If so, can you find which amino acids of this receptor (putative drug target) correspond to the extracellular domain (s) of the protein?

*coursebook: pages 30-34

Alternative ways to access the data



Alternative ways to access the data



- Paste the URL in a location bar in a browser
- Use the terminal window (e.g. with CURL)
- Use our clients (i.e. R and Python)

https://www.targetvalidation.org/documentation/api

Extra hands-on exercises

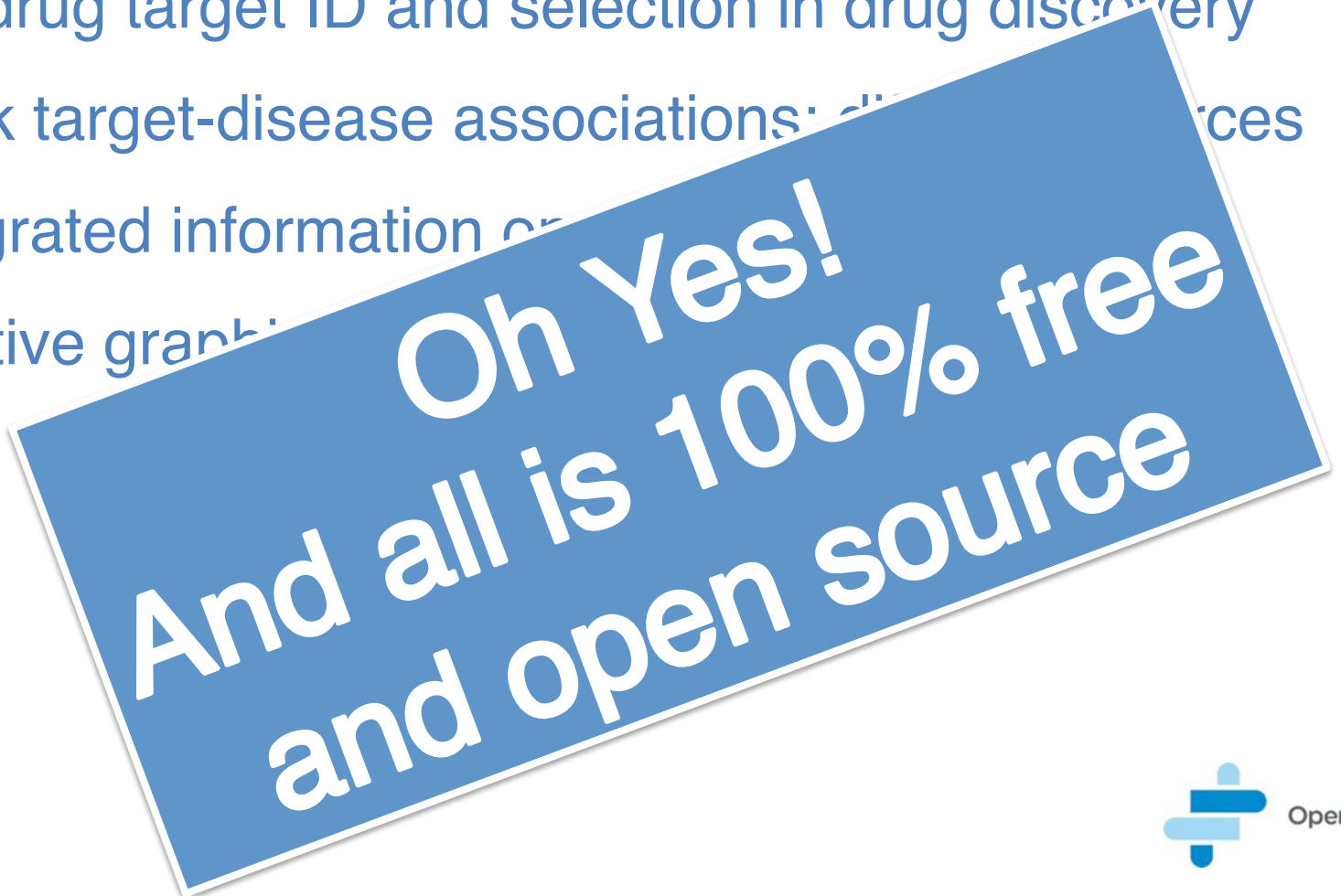
Pages 35-36

Open Targets

Wrap up

Open Targets Platform:

- For drug target ID and selection in drug discovery
- Rank target-disease associations: di… …rces
- Integrated information o…
- Intuitive graph…



Open Targets

Currently: Integration of existing data

Public Databases and Pipelines

ChEMBL
UniProt
Europe PMC
REACTOME
IntAct
intOgen Integrative Onco Genomics
G2P

E
T
D

Open Targets Platform

Open Targets experimental data: NEW
Physiologically relevant and at scale

Oncology

Immunology

Neurodegeneration

Generated as we speak

Open Targets

We support decision-making

Which targets are associated with a disease?

Can I find out about the mechanisms of the disease?

Are there FDA drugs for this association?

What else can I find out about my drug target?

Open Targets

Breakthrough Articles from NUCLEIC ACIDS RESEARCH

http://www.narbreakthrough.com/

How to cite us

Published online 8 December 2016

Nucleic Acids Research, 2017, Vol. 45, Database issue ***D985–D994***
doi: 10.1093/nar/gkw1055

Open Targets: a platform for therapeutic target identification and validation

Gautier Koscielny[1,2,*], Peter An[1,3], Denise Carvalho-Silva[1,4], Jennifer A. Cham[1,4], Luca Fumis[1,4], Rippa Gasparyan[1,3], Samiul Hasan[1,2], Nikiforos Karamanis[1,4], Michael Maguire[1,4], Eliseo Papa[1,3], Andrea Pierleoni[1,4], Miguel Pignatelli[1,4], Theo Platt[1,3], Francis Rowland[1,4], Priyanka Wankar[1,3], A. Patrícia Bento[1,4], Tony Burdett[1,4], Antonio Fabregat[1,4], Simon Forbes[1,5], Anna Gaulton[1,4], Cristina Yenyxe Gonzalez[1,4], Henning Hermjakob[1,4,6], Anne Hersey[1,4], Steven Jupe[1,4], Şenay Kafkas[1,4], Maria Keays[1,4], Catherine Leroy[1,4], Francisco-Javier Lopez[1,4], Maria Paula Magarinos[1,4], James Malone[1,4], Johanna McEntyre[1,4], Alfonso Munoz-Pomer Fuentes[1,4], Claire O'Donovan[1,4], Irene Papatheodorou[1,4], Helen Parkinson[1,4], Barbara Palka[1,4], Justin Paschall[1,4], Robert Petryszak[1,4], Naruemon Pratanwanich[1,4], Sirarat Sarntivijal[1,4], Gary Saunders[1,4], Konstantinos Sidiropoulos[1,4], Thomas Smith[1,4], Zbyslaw Sondka[1,5], Oliver Stegle[1,4], Y. Amy Tang[1,4], Edward Turner[1,4], Brendan Vaughan[1,4], Olga Vrousgou[1,4], Xavier Watkins[1,4], Maria-Jesus Martin[1,4], Philippe Sanseau[1,2], Jessica Vamathevan[4], Ewan Birney[1,4], Jeffrey Barrett[1,4,5] and Ian Dunham[1,4,*]

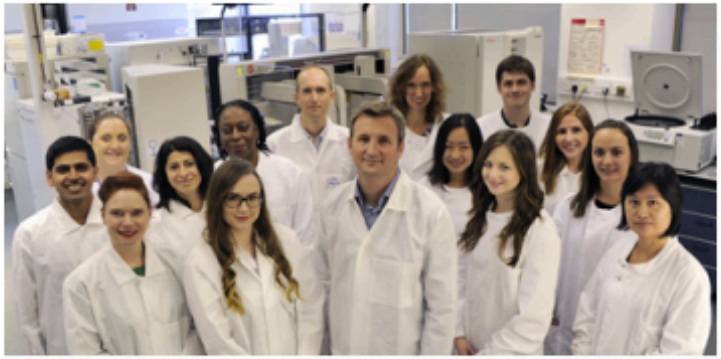
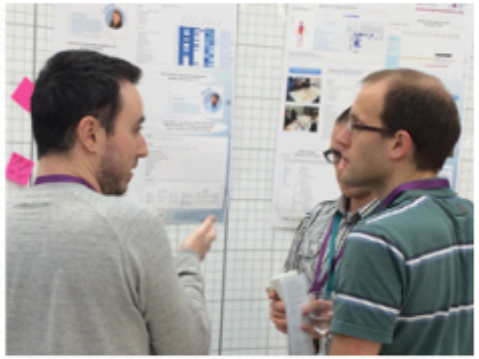
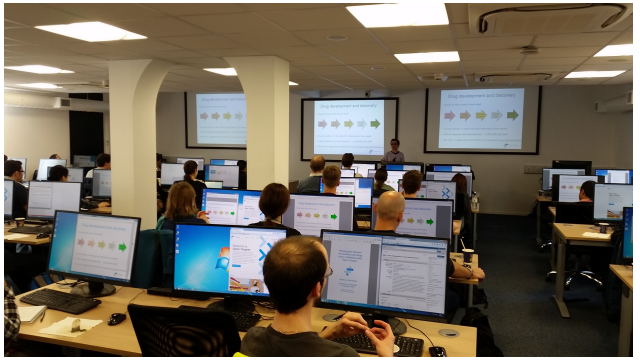
[1]Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [2]GSK, Medicines Research Center, Gunnels Wood Road, Stevenage, SG1 2NY, UK, [3]Biogen, Cambridge, MA 02142, USA, [4]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [5]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and [6]National Center for Protein Research, No. 38, Life Science Park Road, Changping District, 102206 Beijing, China

Received August 19, 2016; Revised October 19, 2016; Editorial Decision October 20, 2016; Accepted November 03, 2016

http://nar.oxfordjournals.org/content/45/D1/D985.long

Acknowledgements



support@targetvalidation.org

Open Targets

Feedback

http://j.tinyurl.com/aruk-030417

Open Targets

Get in touch


@targetvalidate


support@targetvalidation.org


www.facebook.com/OpenTargets/

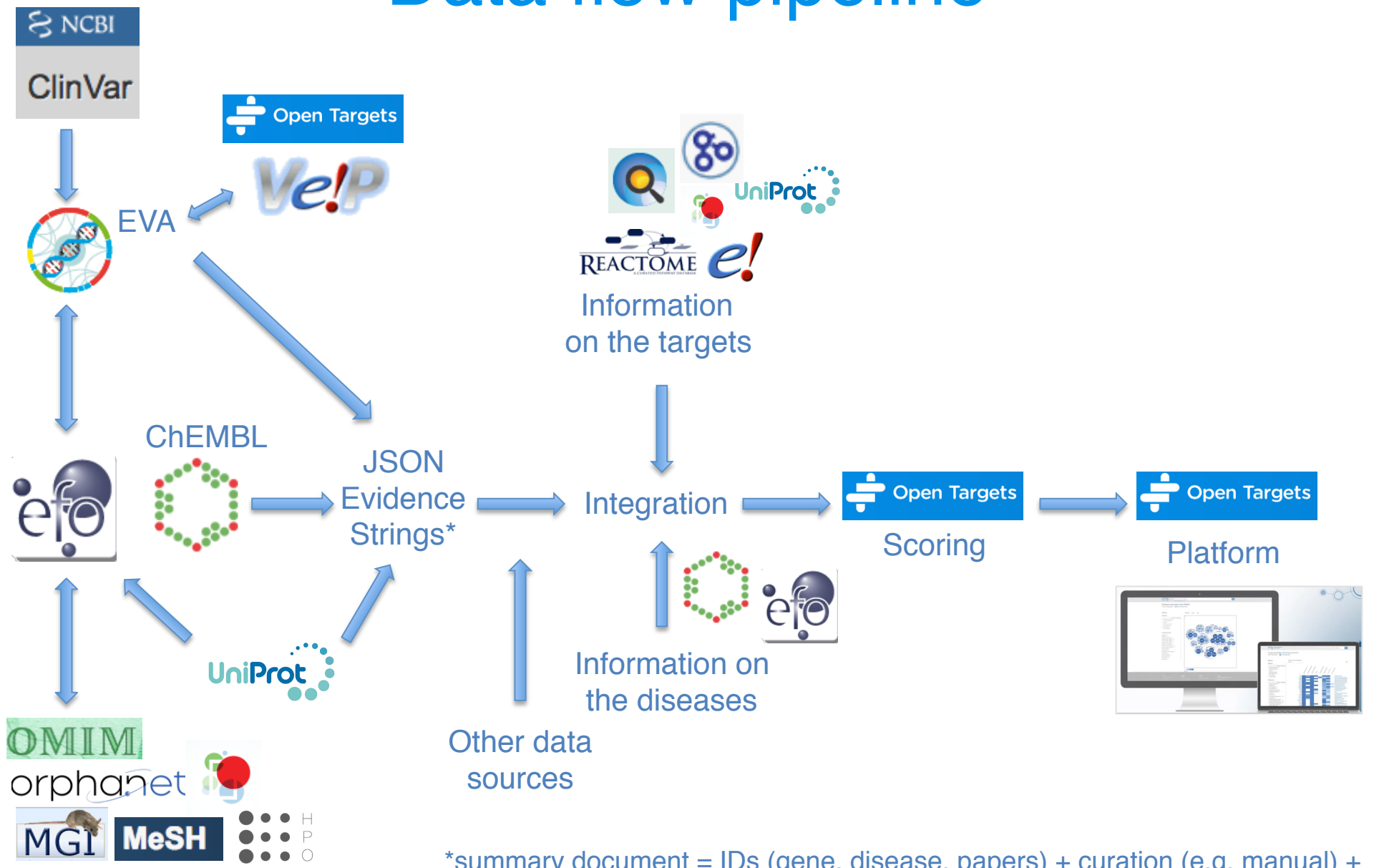

blog.opentargets.org/


http://tinyurl.com/opentargets-in

Open Targets


Open Targets

Data flow pipeline



*summary document = IDs (gene, disease, papers) + curation (e.g. manual) + evidence + source + stats for the score