

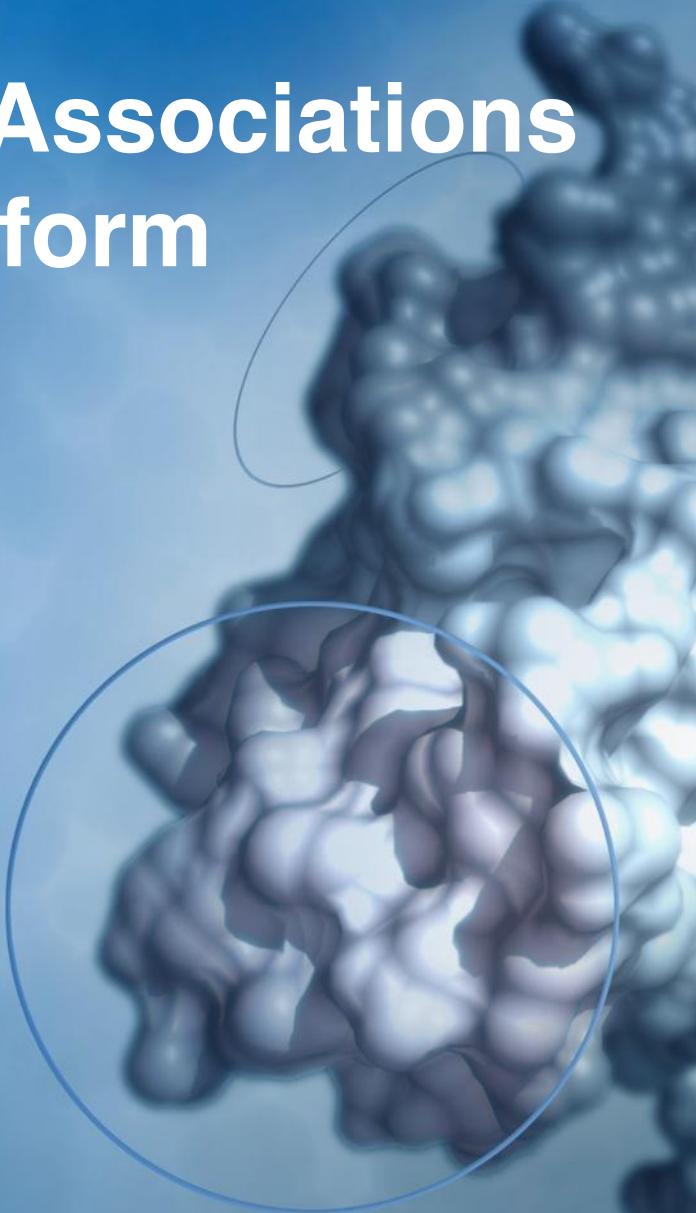
# Mining Gene and Disease Associations with the Open Targets Platform

Drug Discovery Unit  
The University of Dundee

**Denise Carvalho-Silva, PhD**  
Wellcome Genome Campus, United Kingdom  
Open Targets Consortium  
Core Bioinformatics team



Open Targets



# Materials

<https://github.com/deniseOme/training>



DDU\_presentation

DDU\_coursebook

# Course's objectives

What is the Open Targets Platform?

How does Open Targets associate targets with diseases?



How to navigate the Open Targets Platform?

How to connect with the Open Targets team

# Outline

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

# Outline

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

# Two major areas of work in Open Targets

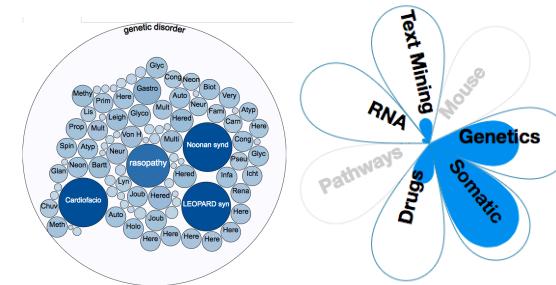
## Experimental projects



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

Concurrent  
[www.opentargets.org/projects](http://www.opentargets.org/projects)

## Core bioinformatics pipelines

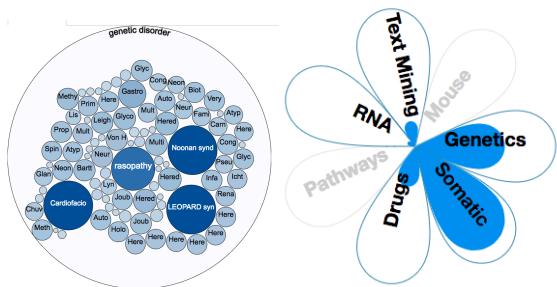


Database for data integration  
Web portal  
REST API and data dumps

# Open Targets Platform\*

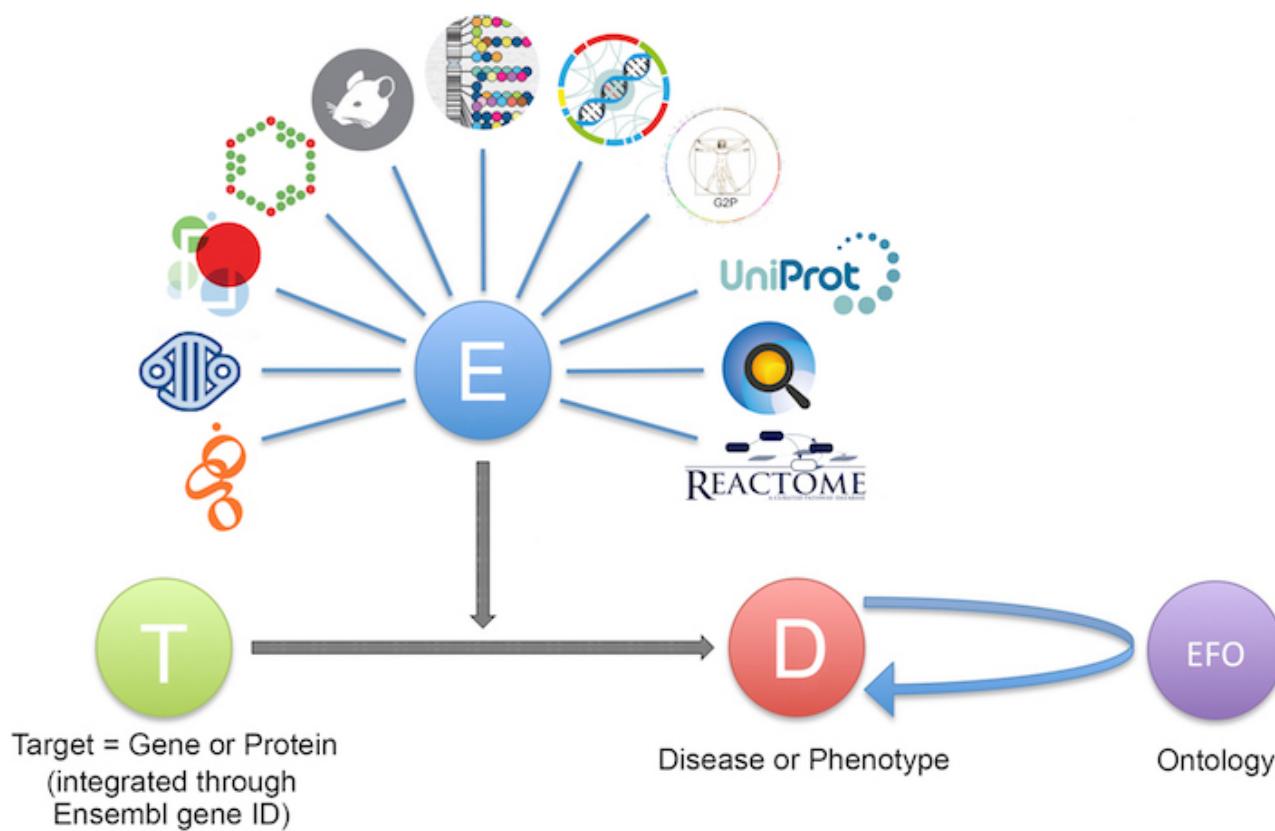
- Developed by the Core Bioinformatics team at EMBL-EBI
- Allow users to identify target and disease associations
- Improvements driven by you

<https://www.targetvalidation.org/>



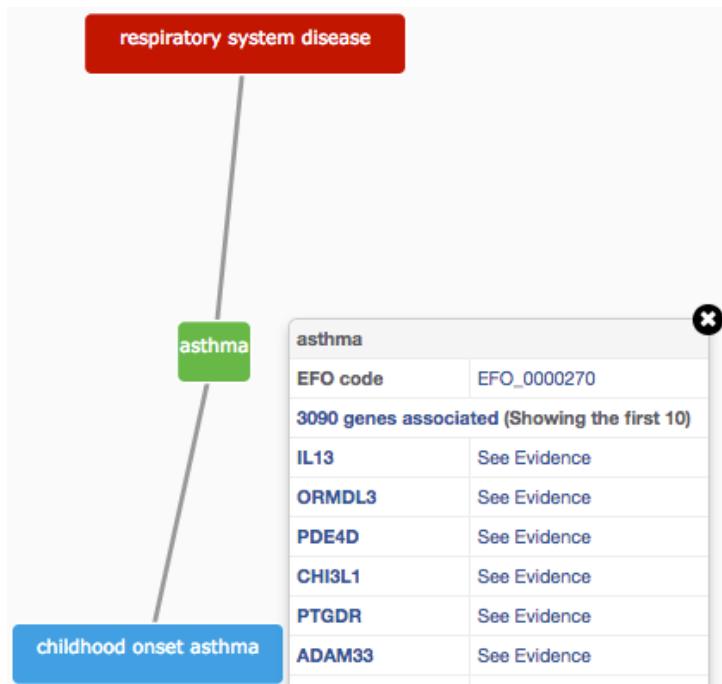
# Open Targets Platform

## Evidence model for target and disease associations



# Experimental Factor Ontology\* (EFO)

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



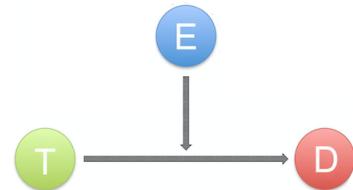
controlled vocabulary  
+  
hierarchy (relationship)

\* <https://www.ebi.ac.uk/efo/>

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

# Evidence from publicly available data

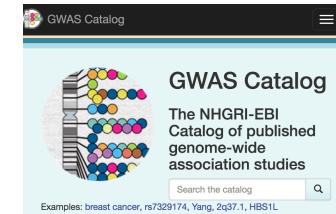
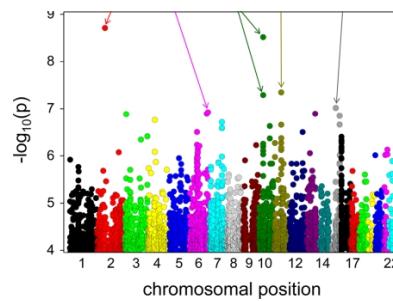
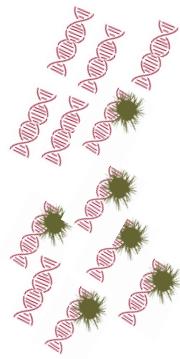
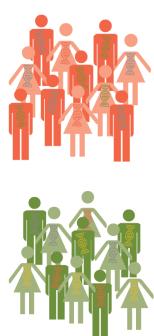
- Similar data sources are grouped into data types



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
<b>Your favourite data?</b>	<b>Let us know!</b>

# 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 in patients



EMBL-EBI train online

# Data sources: EVA

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

The screenshot shows the European Variation Archive (EVA) website. The top navigation bar includes links for Home, Submit Data, Study Browser, Variant Browser, Clinical Browser (which is highlighted in dark blue), GA4GH, API, FAQ, and Feedback. Below the navigation is a search bar with a magnifying glass icon and a "Filter" button. The main content area is titled "ClinVar Browser" with an information icon. It displays a table of results with the following columns: ... (ellipsis), Position, Affecte... (with an info icon), A..., Most Severe Consequence..., Trait, Clinical Significance, and ClinVar ... (with an ellipsis). The table contains 10 rows of data, each corresponding to a variant entry from page 1 of 96. The first row is highlighted in light green.

...	Posi...	Affecte... i	A...	Most Severe Consequence...	Trait	Clinical Significance	ClinVar ...
2	480...	MSH6	T/G	upstream_gen...	Lynch synd...	Benign	RCV000...
2	480...	MSH6	G/A	upstream_gen...	Lynch synd...	Benign	RCV000...
2	480...	MSH6	C/T	upstream_gen...	Lynch synd...	Benign	RCV000...
2	480...	MSH6	C/T	upstream_gen...	Lynch synd...	Benign	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Lynch synd...	Uncertain s...	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...



# Data sources: Gene2Phenotype

Gene2Phenotype

Downloads

Search panel ALL for:  Search

For example: *CRYBA1, ZEB2, TBX1, CHANARIN-DORFMAN SYNDROME* or *MITOCHONDRIAL COMPLEX III DEFICIENCY, NUCLEAR TYPE 1*

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

# Data sources: The Cancer Gene Census

Census

Breakdown

Abbreviations

*The cancer Gene Census is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. The original census and analysis was published in [Nature Reviews Cancer](#) and supplemental analysis information related to the paper is also available.*

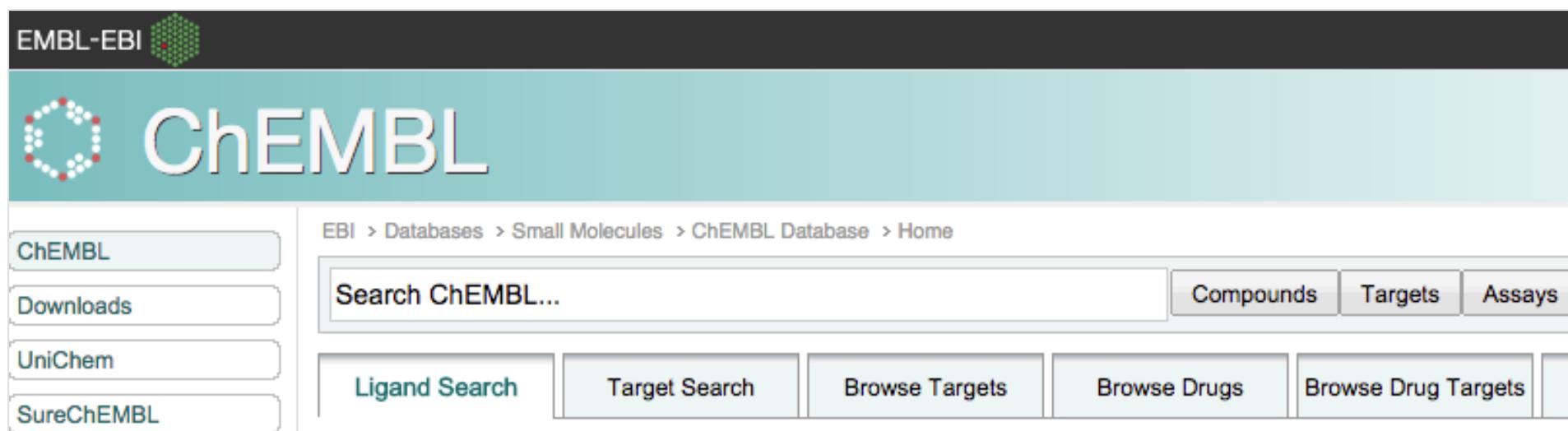
- Genes with mutations causally implicated in cancer
- Gene associated with a cancer plus other cancers associated with that gene

# Data sources: IntOGen

The screenshot shows the IntOGen website. At the top is an orange navigation bar with the IntOGen logo on the left, followed by links for Search, Downloads, Analysis, About, and Sign In. Below the bar is the main content area featuring the IntOGen logo (an orange stylized 'i' icon) and the text "Integrative Onco Genomics".

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

# Data sources: ChEMBL



The screenshot shows the ChEMBL database homepage. At the top left is the EMBL-EBI logo. The main header features the ChEMBL logo (a stylized molecule icon) and the word "ChEMBL". Below the header is a navigation bar with links to "ChEMBL", "Downloads", "UniChem", and "SureChEMBL". To the right of the navigation bar is a search bar containing the placeholder "Search ChEMBL...". Further to the right are three buttons: "Compounds", "Targets", and "Assays". Below the search bar are five buttons: "Ligand Search", "Target Search", "Browse Targets", "Browse Drugs", and "Browse Drug Targets". The page also includes a breadcrumb navigation path: "EBI > Databases > Small Molecules > ChEMBL Database > Home".

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

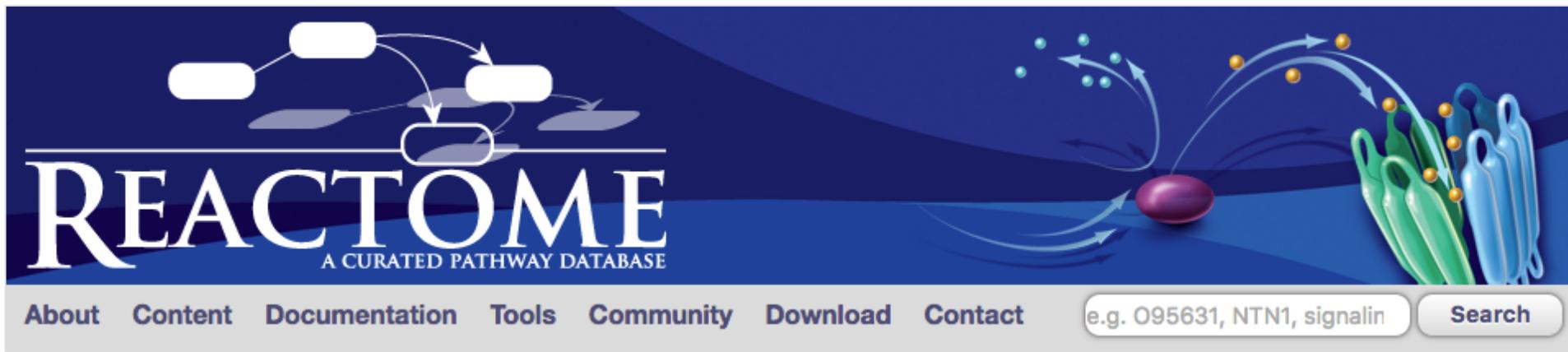


EMBL-EBI train online



Open Targets

# Data sources: Reactome

The image shows the Reactome homepage. At the top, there is a dark blue header with the Reactome logo, which features the word "REACTOME" in large white letters with a horizontal line through it, and "A CURATED PATHWAY DATABASE" below it. To the right of the logo is a decorative graphic of molecular structures and arrows. Below the header is a navigation bar with links: "About", "Content", "Documentation", "Tools", "Community", "Download", and "Contact". To the right of the navigation bar is a search bar containing the placeholder text "e.g. O95631, NTN1, signalin" and a "Search" button.

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



EMBL-EBI train online



Open Targets

# Data sources: Expression Atlas

The screenshot shows the Expression Atlas homepage. At the top, there is a dark header bar with the EMBL-EBI logo on the left and navigation links for Services, Research, Training, and About us on the right. Below the header, the main title "Expression Atlas" is displayed, featuring a magnifying glass icon next to the word "Expression". To the right of the title is a search bar with the placeholder "Enter gene query..." and a "Search" button. Below the search bar, there is an example query: "ASPM, Apoptosis, ENSMUSG00000021789, zinc finger". A navigation menu bar below the title includes links for Home, Release notes, FAQ, Download, Help, Licence, and About. On the far right of this menu bar is a "Feedback" link with a speech bubble icon.

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

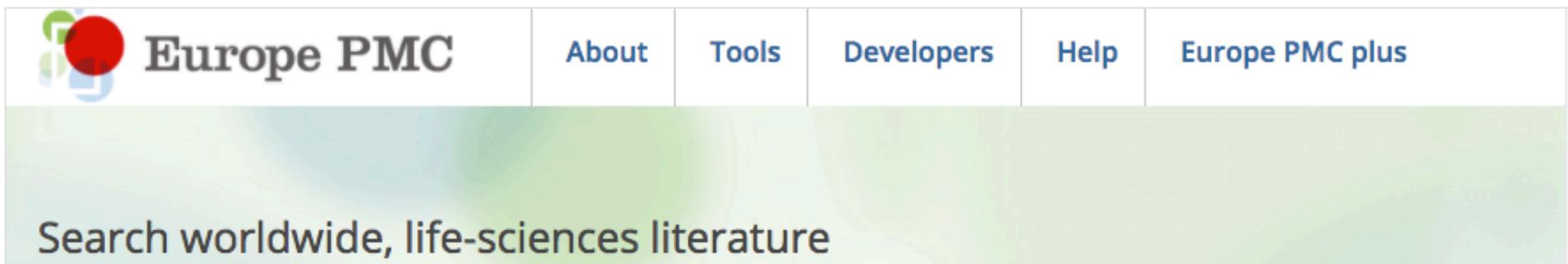


EMBL-EBI train online



Open Targets

# Data sources: Europe PMC



A screenshot of the Europe PMC website. At the top, there is a navigation bar with links for "About", "Tools", "Developers", "Help", and "Europe PMC plus". To the left of the navigation bar is the Europe PMC logo, which consists of three stylized green and blue overlapping shapes next to the text "Europe PMC". Below the navigation bar is a large search bar with the placeholder text "Search worldwide, life-sciences literature".

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



EMBL-EBI train online



Open Targets

# Data sources: PhenoDigm

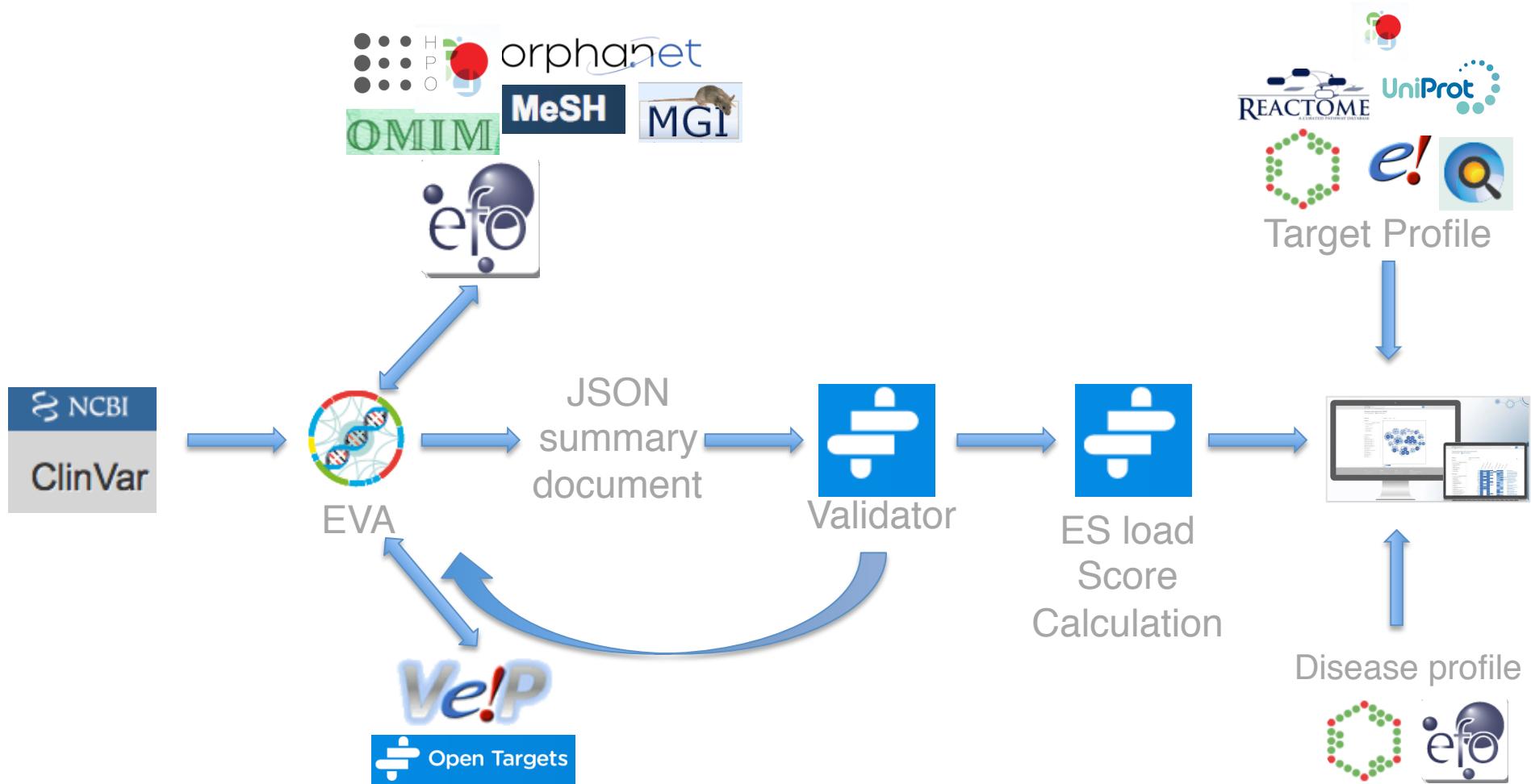
The screenshot shows the homepage of the PhenoDigm website. At the top, there is a dark header bar with the Wellcome Trust Sanger Institute logo on the left. To the right of the logo is a blue navigation bar with the following links: "ABOUT" (with a dropdown arrow), "Who we are", "Careers", "Study", "Sex in Science", "Groups", and "Campus". On the far right of the blue bar is a magnifying glass icon for search. Below the header, the main title "Welcome to PhenoDigm (PHENOtype comparisons for Disease and Gene Models)" is displayed in large, bold, black font. Underneath the title, there is a horizontal menu bar with three items: "Diseases" (which is highlighted in blue), "Tissue phenotype associations", and "Secondary phenotypes".

## Welcome to PhenoDigm (PHENOtype comparisons for Disease and Gene Models)

Diseases Tissue phenotype associations Secondary phenotypes

- Semantic approach to associate mouse models with diseases

# Data flow pipeline

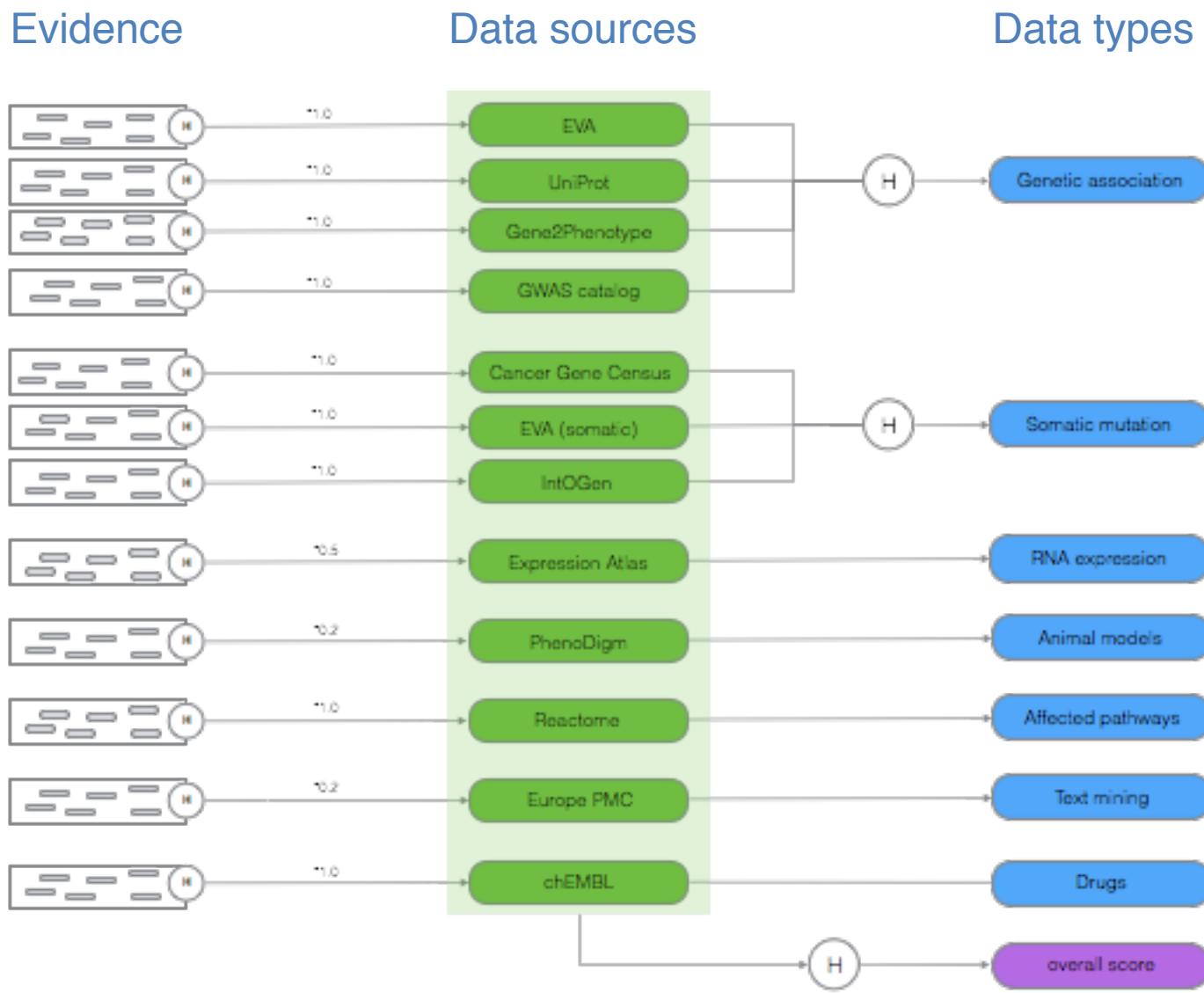


\* JSON summary document = IDs (gene, disease, papers) + curation (e.g. manual) + evidence + source + stats for the score

# JSON summary document

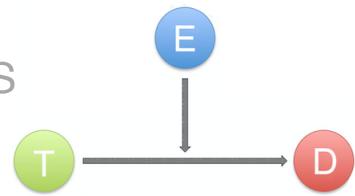
```
ArrayExpress Experiment overview"}], "experiment_overview": "Transcription profiling of human intestinal epithelium in patients suffering from inflammatory bowel disease.", "access_level": "public", "id": "866f4a506d5a01ebcea8bd84bb5485c8", "unique_association_fields": {"comparison_name": "'active ulcerative colitis' vs 'normal'", "geneID": "http://identifiers.org/ensembl/ENSG00000073605", "study_id": "http://identifiers.org/gxa.expt/E-MEXP-2083"}, "sourceID": "expression_atlas", "target": {"gene_info": {"symbol": "GSDMB", "geneid": "ENSG00000073605", "name": "gasdermin B"}, "id": "ENSG00000073605", "activity": "increased_transcript_level", "target_type": "transcript_evidence"}, "disease": {"id": "EFO_0000729", "efo_info": {"path": [[{"id": "EFO_0000405", "label": "ulcerative colitis"}, {"id": "EFO_0003767", "label": "therapeutic area"}, {"id": "EFO_0000729", "label": "immune system disease"}, {"id": "EFO_0000540", "label": "digestive system disease"}], "label": "ulcerative colitis", "therapeutic_area": {"labels": ["immune system disease", "digestive system disease"]}, "codes": [{"id": "EFO_0000405", "label": "ulcerative colitis"}, {"id": "EFO_0000540", "label": "digestive system disease"}]}, "biosample": {"name": "intestinal epithelial cell", "id": "http://purl.obolibrary.org/obo/CL_0002563"}, "data_release": "16.12", "type": "rna_expression", "scores": {"association_score": 0.014397535094125722}}  
{"validated_against_schema_version": "1.2.3", "evidence": {"confidence_level": "low", "log2_fold_change": {"value": 1, "percentile_rank": 97}, "unique_experiment_reference": "STUDYID_E-GEOID-48258", "resource_score": {"value": 0.0328, "type": "pvalue", "method": {"description": "Moderated t-statistics computed with limma version 3.16.8. Independent filtering performed using genefilter version 1.46.1, with gene variances as filter statistic. p-values adjusted using Benjamini & Hochberg (1995) FDR correction. By default, significant differential expression called if adjusted p <= 0.05."}}, "date_asserted": "2015-07-20T22:01:23Z", "reference_replicates_n": 3, "evidence_codes_info": [{"eco_id": "ECO_0000356", "label": "differential gene expression evidence from microarray experiment"}], "test_replicates_n": 3, "is_associated": true, "provenance_type": {"database": {"version": "18-11-2016", "id": "Expression_Atlas"}, "evidence_codes": [{"id": "ECO_0000356"}], "reference_sample": "control", "comparison_name": "'fludarabine, 10 micromolar' vs 'control'", "test_sample": "fludarabine, 10 micromolar", "url": [{"url": "http://www.ebi.ac.uk/gxa/experiments/E-GEOID-48258?geneQuery=ENSG00000179889", "nice_name": "Gene expression in Expression Atlas"}, {"url": "http://www.ebi.ac.uk/gxa/genes/ENSG00000179889", "nice_name": "Basel"}]}
```

# Score approach and aggregation



Evidence score e.g.  $f * S * C$  (frequency x Severity x Confidence) for GWAS

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



# Outline

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

# Demo 1: Evidence for a T-D association

What is the evidence for the association between *CD86* and multiple sclerosis?



Open Targets Platform ≡ Q

Evidence for CD86 in multiple sclerosis

**CD86**  
CD86 molecule  
Synonyms: B7.2, B7-2, CD28LG2

**multiple sclerosis**  
Synonyms: MS (Multiple Sclerosis), MS, MULTIPLE SCLEROSIS ACUTE FULMINATING, Disseminated Sclerosis, Sclerosis...

Target profile page Disease profile page

A large red arrow points downwards from the evidence diagram towards the URL at the bottom of the slide.

[https://www.targetvalidation.org/evidence/ENSG00000114013/EFO\\_0003885](https://www.targetvalidation.org/evidence/ENSG00000114013/EFO_0003885)

# Choose your favourite internet browser

Supported ones: Internet Explorer 11 (and above), Chrome, Firefox and Safari



# Demo 2: Several targets to pursue



We have a list of 26 possible targets for IBD (inflammatory bowel disease).

How can we get all the data in Open Targets for all of them using the website?

<https://www.targetvalidation.org/batch-search>



# Hands-on exercises

Pages 26-29

# Hands-on exercises

Pages 26-29

Not enough time left today?

Email us with your questions



[support@targetvalidation.org](mailto:support@targetvalidation.org)

# Wrap up

## Open Targets Platform:

For drug target ID and selection in drug discovery

Rank target-disease associations: disease prioritization

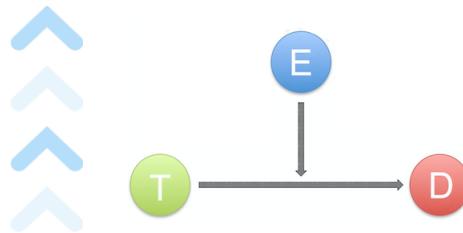
Integrated information on targets and diseases

Intuitive graphical interface

Oh Yes!  
And all is 100% free  
and open source

# Currently: Integration of existing data

## Public Databases and Pipelines



Open Targets experimental data: NEW  
Physiologically relevant and at scale

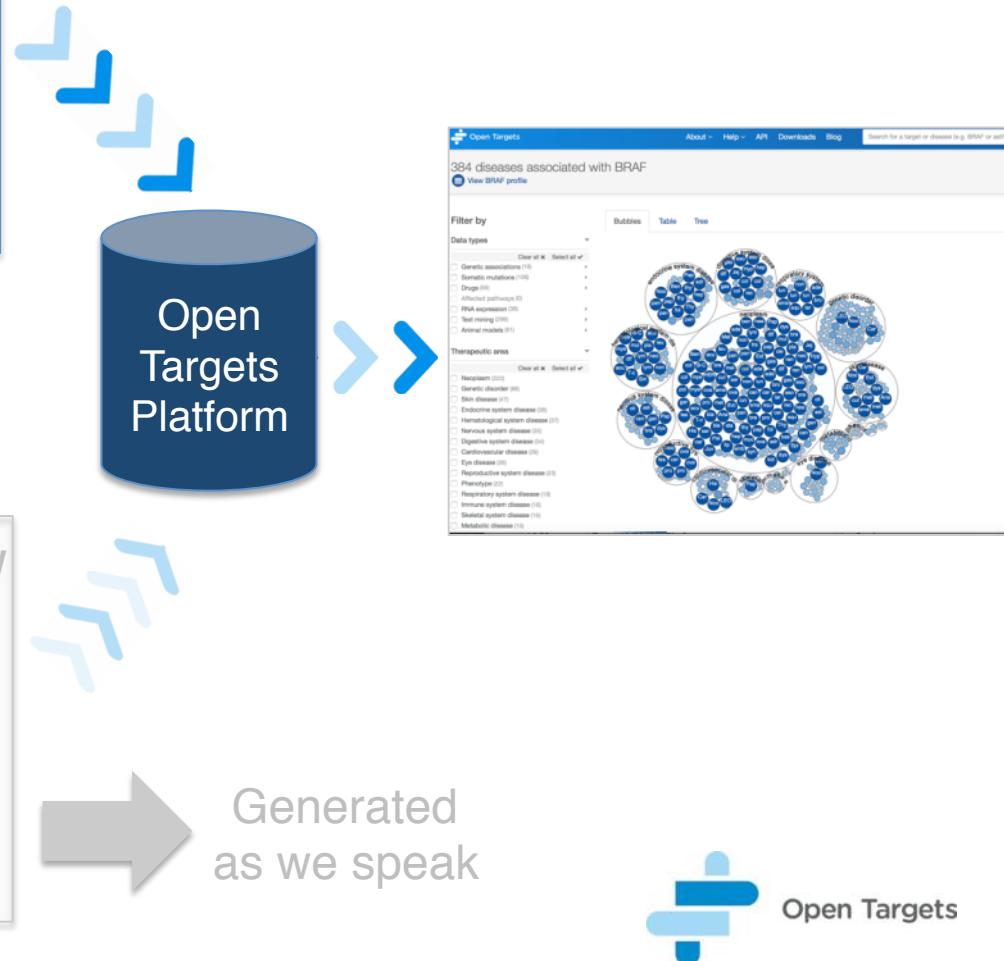
Oncology



Immunology



Neurodegeneration



Open Targets

# We support decision-making

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

# Alternative ways to access the data

The screenshot shows a web browser window with the URL <https://www.targetvalidation.org/download> in the address bar. The page itself has a blue header with the Open Targets Platform logo and navigation icons. The main content area is titled "Data Download" and contains text explaining that all data from targetvalidation.org is available for download as compressed JSON files. It describes the availability of associations and evidence objects via API methods. Below this, a section titled "2017 Feb (Latest)" lists two download links: "Association objects (2016-12-09, 215MB, md5sum)" and "Evidence objects (2016-12-09, 4.35Gb, md5sum)".

All data from targetvalidation.org is available for download as compressed JSON files.

We provide downloads of all associations between target and disease calculated by the platform, as well as all the evidence used in calculating each associations. These are the same objects returned by the corresponding [/public/associations](#) and [/public/evidence](#) API methods. Head to the API documentation for further details.

**2017 Feb (Latest)**

- Association objects (2016-12-09, 215MB, md5sum)
- Evidence objects (2016-12-09, 4.35Gb, md5sum)

# Alternative ways to access the data



public : Publicly supported stable API.

Open/Hide | List operations | Expand operations

GET /public/evidence

POST /public/evidence

GET /public/evidence/filter

POST /public/evidence/filter

GET /public/association

GET /public/association/filter

POST /public/association/filter

GET /public/search

GET /public/auth/request\_token

GET /public/auth/validate\_token

GET /public/utils/ping

GET /public/utils/version

GET /public/utils/stats

- 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>



Open Targets

# API calls: some examples

GET

/public/search

[http://targetvalidation.org/api/latest/public/search?q=EFO\\_0003767](http://targetvalidation.org/api/latest/public/search?q=EFO_0003767)

GET

/public/association/filter

[http://www.targetvalidation.org/api/latest/public/association/filter?  
target=ENSG00000110324&direct=false&fields=is\\_direct&fields=disease.efo\\_info.lab  
el&size=100](http://www.targetvalidation.org/api/latest/public/association/filter?target=ENSG00000110324&direct=false&fields=is_direct&fields=disease.efo_info.label&size=100)

GET

/public/evidence/filter

[https://targetvalidation.org/api/latest/public/evidence/filter?  
target=ENSG00000141867&disease=EFO\\_0000565&datatype=expression\\_atl  
as&size=100&format=json](https://targetvalidation.org/api/latest/public/evidence/filter?target=ENSG00000141867&disease=EFO_0000565&datatype=expression_atlas&size=100&format=json)

# Outline

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

# Outreach



[support@targetvalidation.org](mailto:support@targetvalidation.org)



<http://tinyurl.com/opentargets-in>



[@targetvalidate](#)



[blog.opentargets.org/](http://blog.opentargets.org/)



[www.facebook.com/OpenTargets/](http://www.facebook.com/OpenTargets/)



Open Targets

# How to cite us

Published online 8 December 2016

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

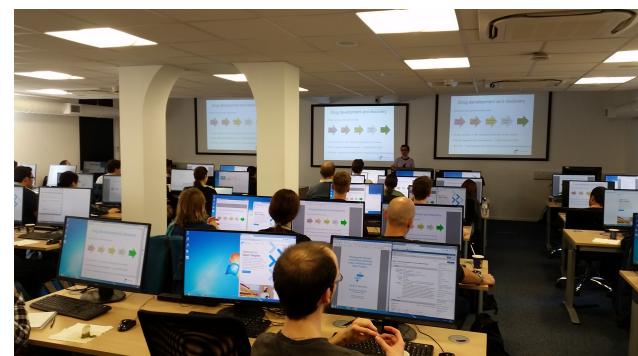
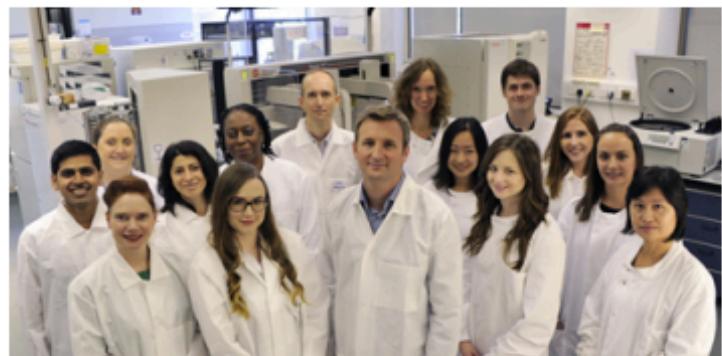
## Open Targets: a platform for therapeutic target identification and validation

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

<sup>1</sup>Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, <sup>2</sup>GSK, Medicines Research Center, Gunnels Wood Road, Stevenage, SG1 2NY, UK, <sup>3</sup>Biogen, Cambridge, MA 02142, USA, <sup>4</sup>European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, <sup>5</sup>Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and <sup>6</sup>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

# Acknowledgements



[support@targetvalidation.org](mailto:support@targetvalidation.org)



# Feedback survey

<http://bit.ly/2q1aXWc>

Extra Extra Extra  
(slides)

# 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

# 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 and controls)

s = predicted functional consequence

c = *p* value reported in the paper

# Aggregating scores across the data

- Using a mathematical function, the harmonic sum\*

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

where  $S_1, S_2, \dots, S_i$  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

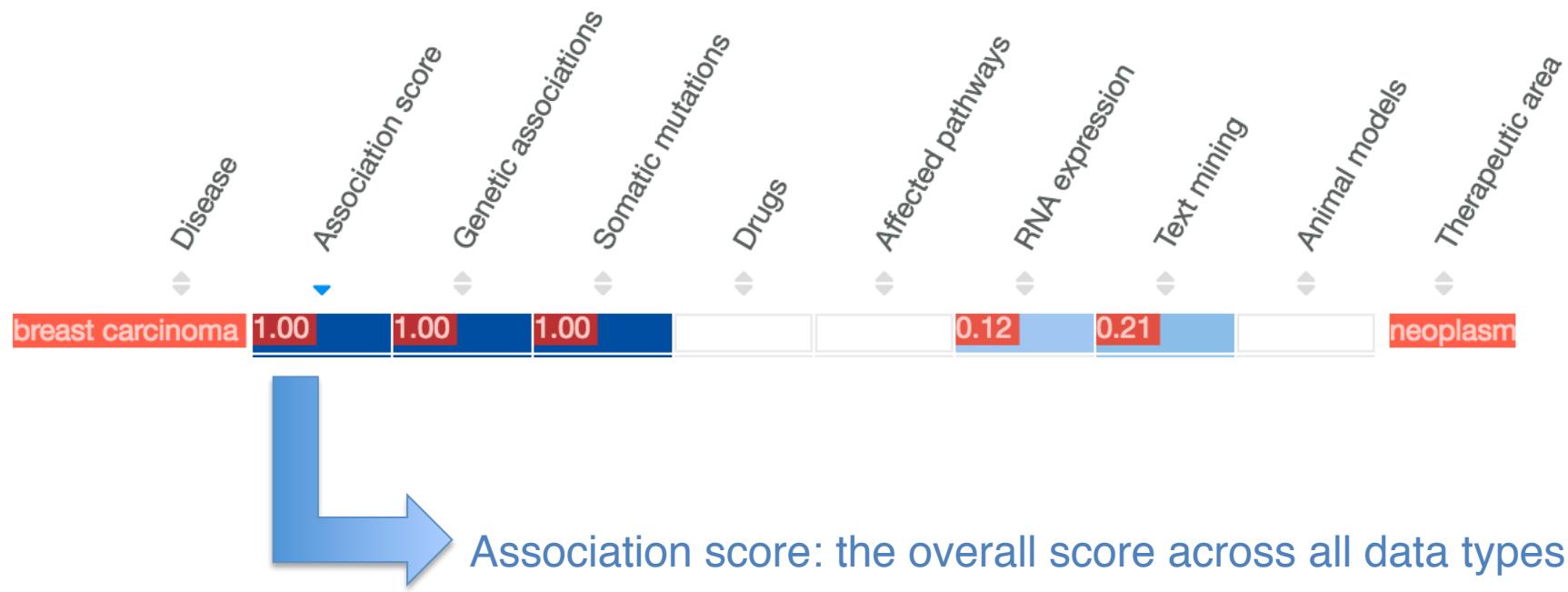
# Disclaimer: score, dos and don'ts

- 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

# 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