

# 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



# 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

# Drug discovery path: timeline

## 1. DISCOVERY



IDEA



### BASIC RESEARCH

The majority of the research at this stage is publicly funded at universities, colleges and independent research institutions in every state.

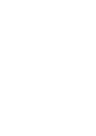
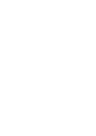
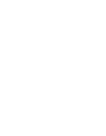
## 2. DEVELOPMENT



### CLINICAL TRIALS

Once a disease target is identified, drugs are designed and tested. Both public and privately funded research are involved.

PHASE I   PHASE II   PHASE III



## 3. DELIVERY



### REGULATORY APPROVAL

Human trials are completed. FDA approval. Industry is responsible for bringing a drug to market. Safety and evaluation continue after approvals.

### PATIENT CARE

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



*Professor Sir  
Mike Stratton  
Director, Sanger Institute*

Can we improve  
target identification?



*Patrick Vallance, President  
Pharmaceuticals R&D  
GlaxoSmithKline*



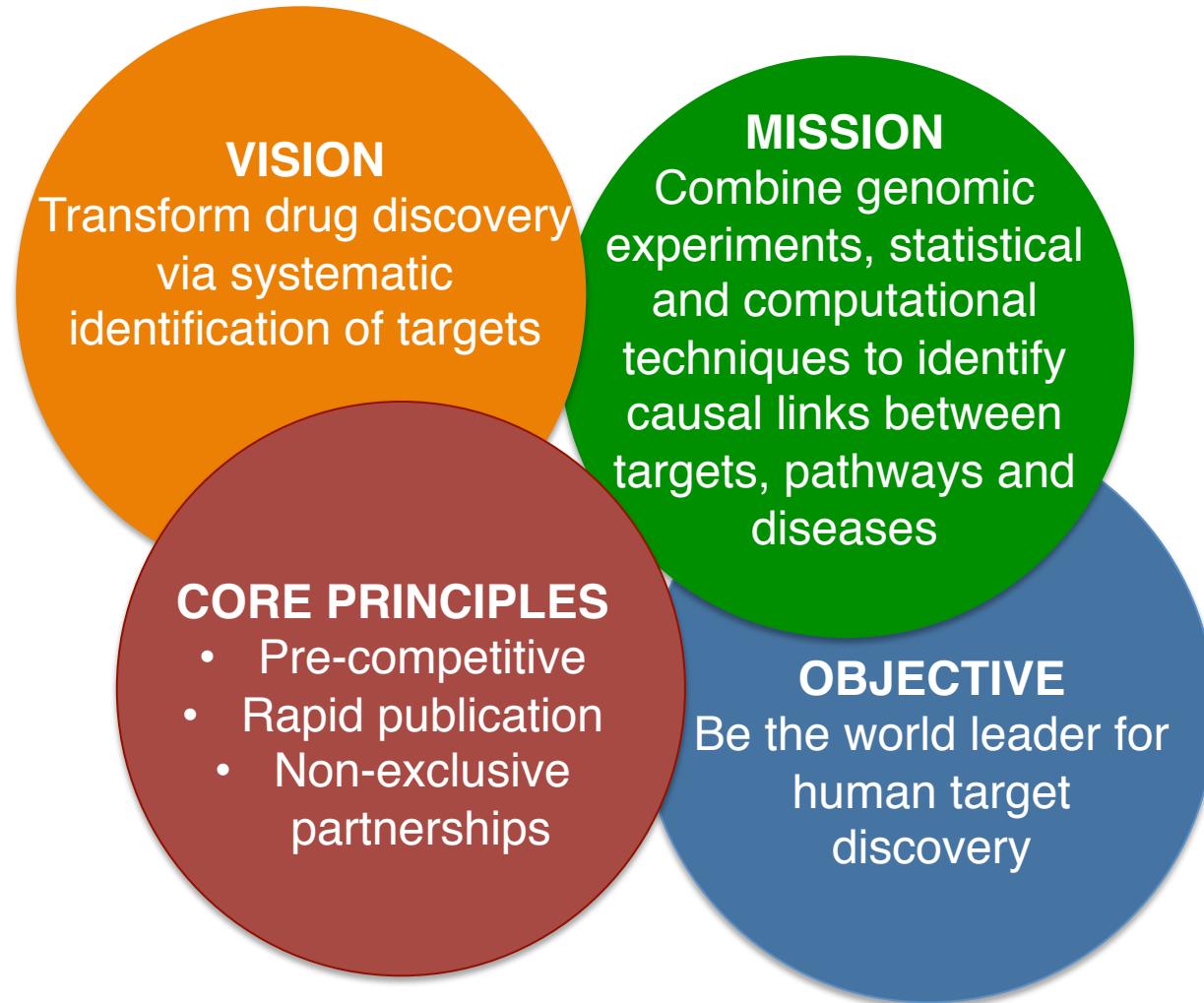
*Professor Dame  
Janet Thornton  
former Director, EMBL-EBI*

Yes, we can!  
And we should.

But one institution  
can not do it alone.



# Open Targets Consortium\*



\* Three founding partners, March 2014



EMBL-EBI



# Who is Open Targets today?

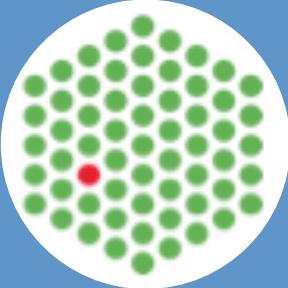
GSK



Expertise in disease biology  
Translational medicine



EMBL-EBI



Experts in life-science data integration and analysis  
Access to vast public domain resources



WTSI



Expertise in the role of genetics in disease  
Extensive experimental capabilities



Biogen\*



Leading in neurodegenerative diseases and innovative hemophilia therapies



\* Biogen joined the consortium in February 2016



Open Targets

# Two major areas of work in Open Targets

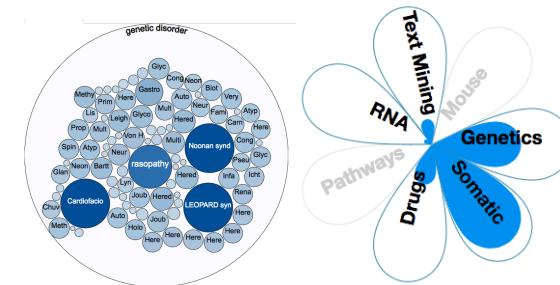
## Experimental projects



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

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

## Core bioinformatics pipelines



Integration of available data  
Web portal  
REST API and data dumps

# Two major areas of work in Open Targets

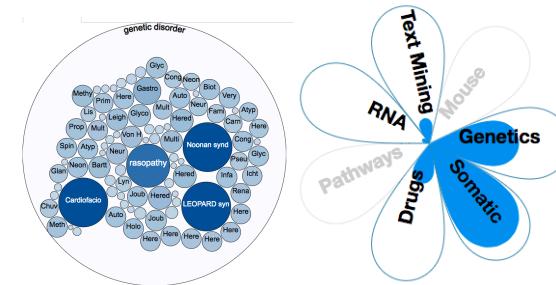
## Experimental projects



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

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

## Core bioinformatics pipelines

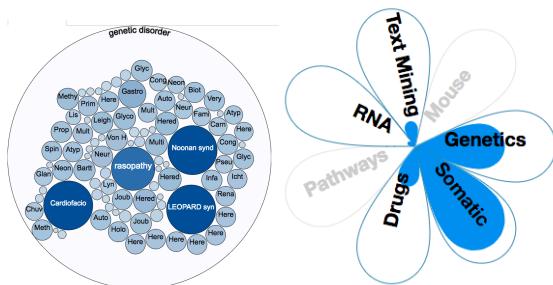


Database for data integration  
Web portal  
REST API and data dumps

# Open Targets Platform\*

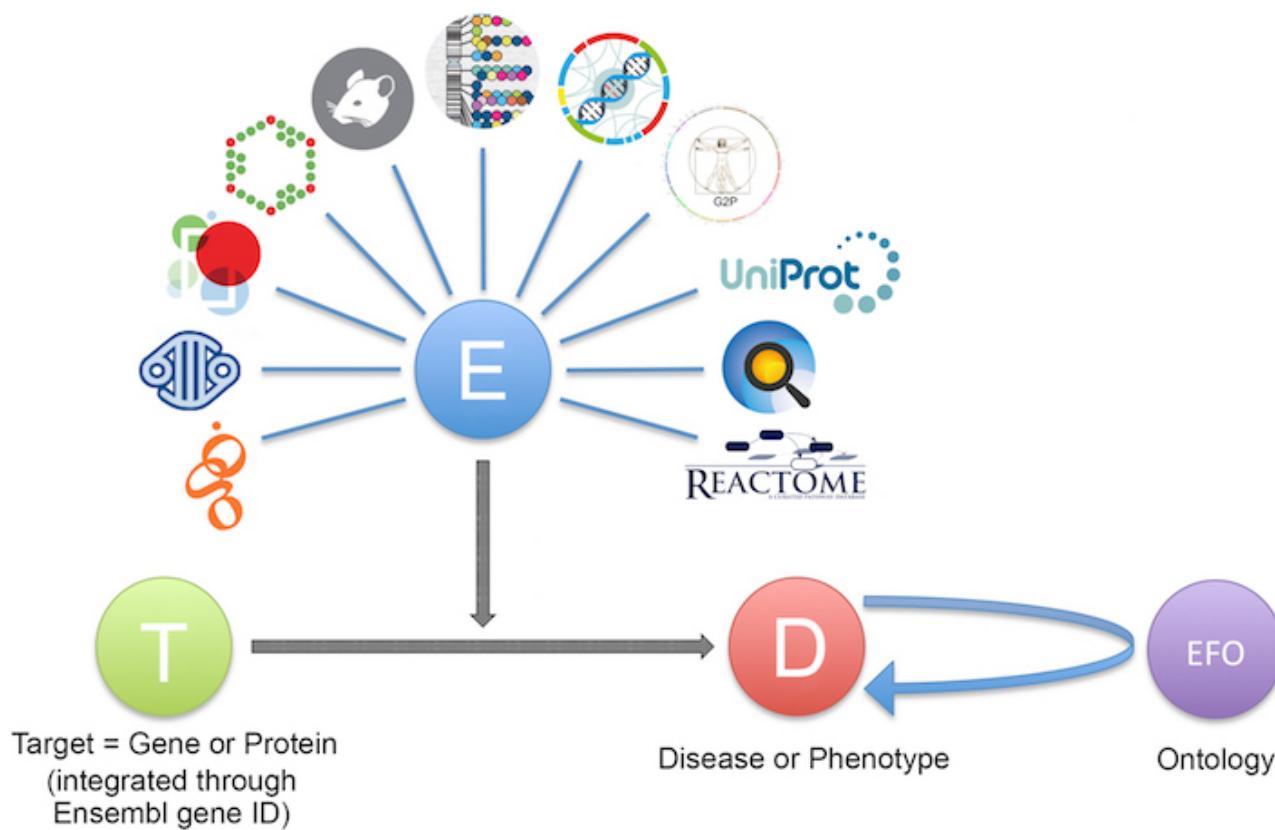
- Developed by the Core Bioinformatics team at EMBL-EBI
- Allow users to identify target–disease associations
- Improvements driven by you

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



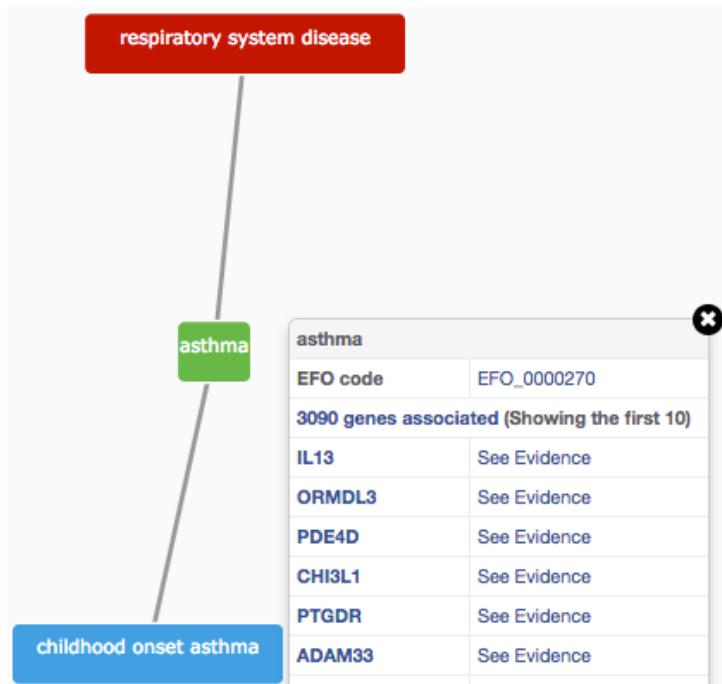
# Open Targets Platform

## Evidence model for target–disease associations



# Experimental Factor Ontology\* (EFO)

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



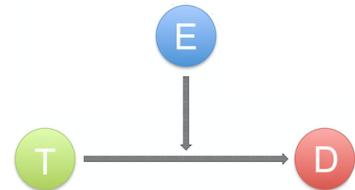
controlled vocabulary  
+  
hierarchy (relationship)

\* <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

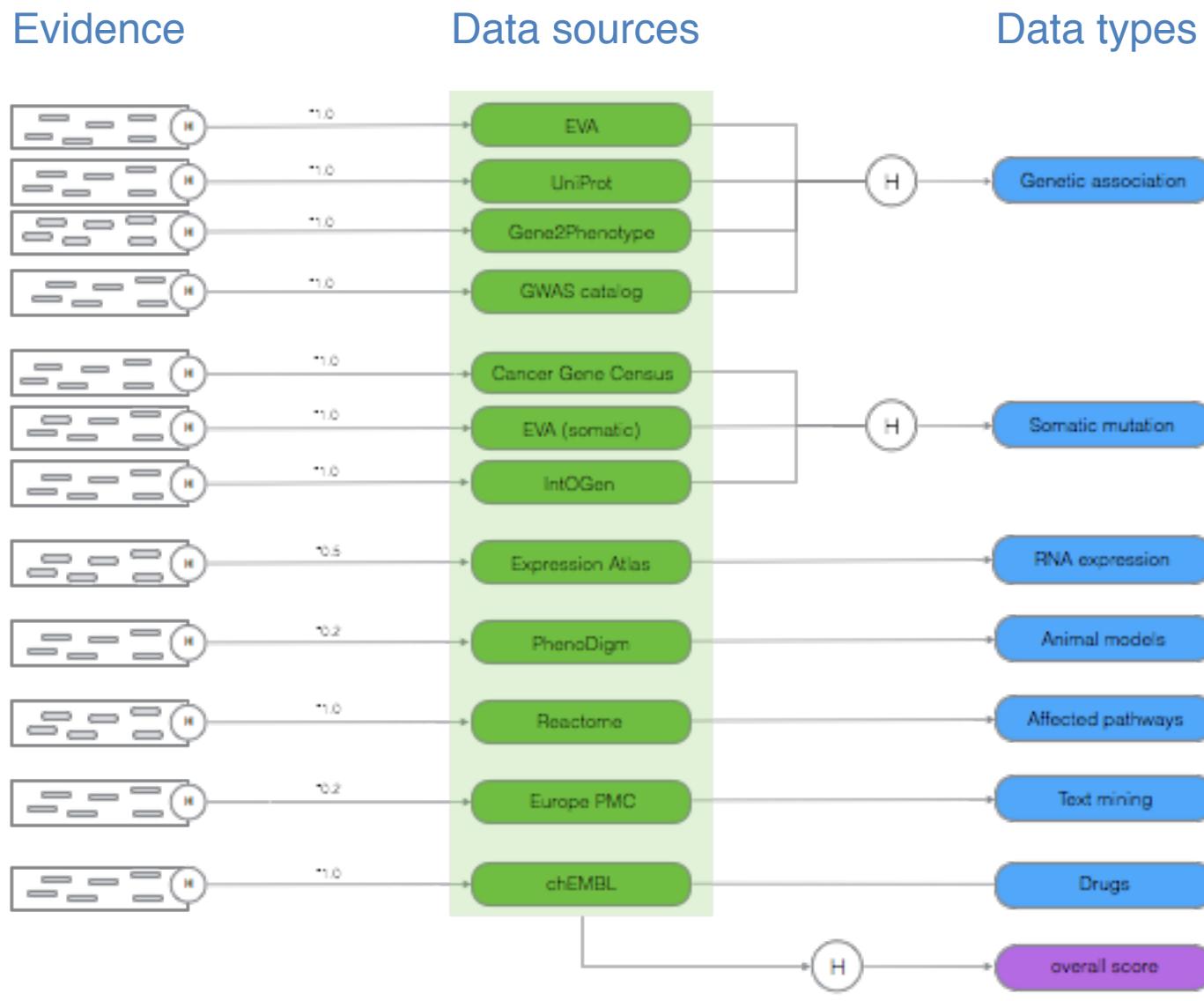


Data sources	Data types
GWAS catalog, UniProt, EVA, G2P	Genetic associations
Cancer Gene Census, EVA, IntOgen	Somatic mutations
Expression Atlas	RNA expression
ChEMBL	Drugs
Reactome	Affected pathways
Europe PMC	Text mining
PhenoDigm	Animal models
<b>Your favourite data?</b>	<b>Let us know!</b>

# How do we associate diseases and phenotypes w/ targets?

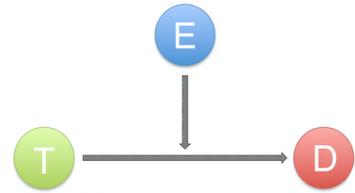
- 1 ChEMBL, UniProt, EVA (w/ ClinVar) curate diseases and phenotypes
- 2 Map disease/phenotypes to an ontology using EFO and HPO terms
- 3 Use genes as proxies for our targets
- 4 Create target-disease evidence JSON objects
- 5 Calculate for each supporting evidence the likelihood of gene A being associated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

# Score approach and aggregation



$$\text{Evidence score} = f * S * C \text{ (frequency x Severity x Confidence)}$$

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



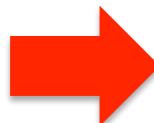
Which targets are associated with a disease?



## Demo 1:

screenshots: next slides  
coursebook: pages 9-15

The screenshot shows the homepage of the Open Targets Platform. At the top, there is a navigation bar with links for Survey, About, Help, API, Downloads, and Blog. Below the navigation is the platform's logo, "Open Targets Platform", and the tagline "Find new targets for drug discovery". A search bar contains the placeholder "Search for a target or disease" with a magnifying glass icon. Below the search bar, there is a "Try:" section with examples: BRAF, PTEN, Asthma, and Inflammatory bowel disease. On the far right, there are "Feedback" and "Follow us" buttons.



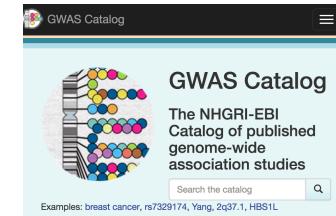
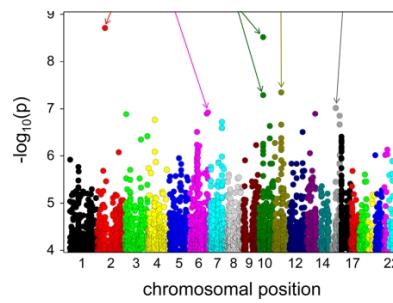
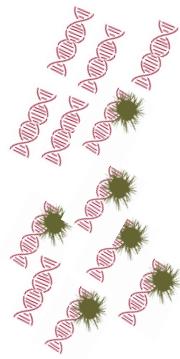
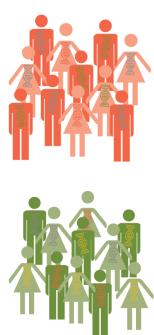
The screenshot shows the search results for "multiple sclero". The search bar at the top contains "multiple sclero". The main result is "multiple sclerosis" with "2697 targets associated". It is categorized as a "Disease". A brief description follows: "An autoimmune disorder mainly affecting young adults and characterized by destruction of myelin in the central nervous system. Pathologic findings include multiple sharply demarcated areas of demyelination throughout the white matter of the central nervous system. Clinical manifestations include vis...". Below this, under "Targets", is "MBP myelin basic protein". Under "Diseases", it lists "relapsing-remitting multiple sclerosis" and "autoimmune disease > multiple sclerosis > relapsing-remitting multiple ...". Another section for "chronic progressive multiple sclerosis" is also shown, with a similar hierarchical path: "autoimmune disease > multiple sclerosis > chronic progressive multiple...".

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



# 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



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 currently selected), GA4GH, API, FAQ, and Feedback. Below the navigation is a search bar with a placeholder 'Search' and a 'Go' button. To the left, there's a 'Filter' section with buttons for 'Reset' and 'Sub...', and dropdown menus for 'Position' (set to GRCh37), 'Assembly' (set to GRCh37), 'Filter By' (set to Chromosomal), and a specific position range '2:48000000-49000000'. The main content area is titled 'ClinVar Browser' and displays a table of results. The table has columns for Position, Affect., Most Severe Consequence, Trait, Clinical Significance, and ClinVar ID. The first few rows show variants for the MSH6 gene at position 2, with various consequence types like upstream\_genic, Lynch syndrome, and Benign.

...	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...



# 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



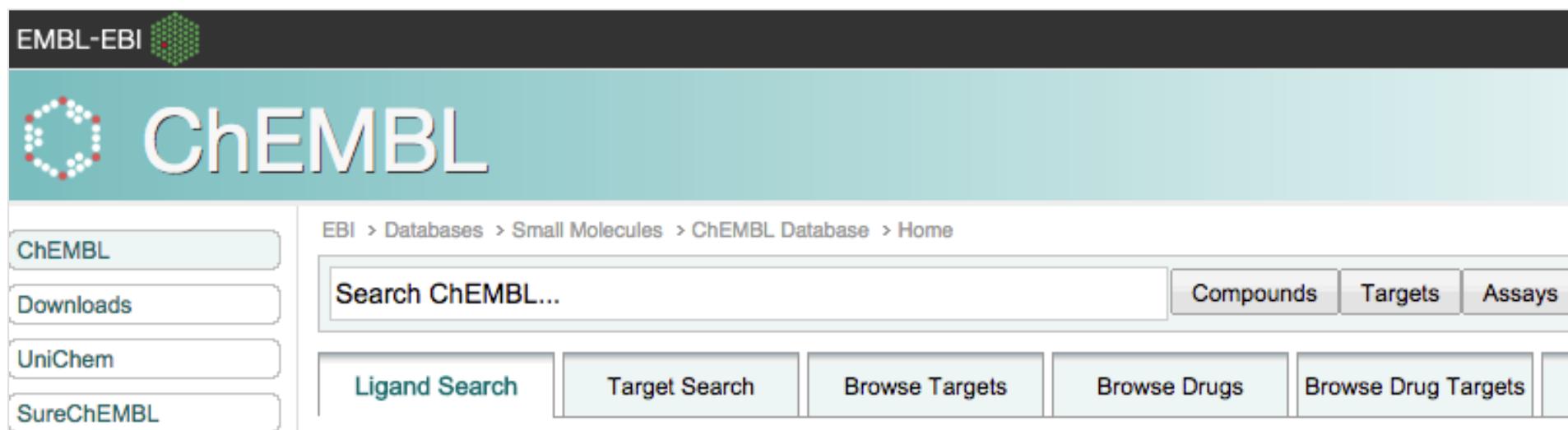
Open Targets

# Data sources: IntOGen

The screenshot shows the homepage of the intOGen website. At the top is a navigation bar with links for Search, Downloads, Analysis, and About. On the far right is a Sign In button. Below the navigation bar is the intOGen logo, which consists of a stylized orange 'i' icon followed by the word 'intOGen' in lowercase. To the right of the logo is the text 'Integrative Onco Genomics' in orange. The main content area below the logo contains a large, faint watermark-like image of a brain.

- 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 circular 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 breadcrumb trail: "EBI > Databases > Small Molecules > ChEMBL Database > Home". Below the breadcrumb is a search bar with the placeholder "Search ChEMBL...". To the right of the search bar are three buttons: "Compounds", "Targets", and "Assays". At the bottom of the page are five buttons: "Ligand Search", "Target Search", "Browse Targets", "Browse Drugs", and "Browse Drug Targets".

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

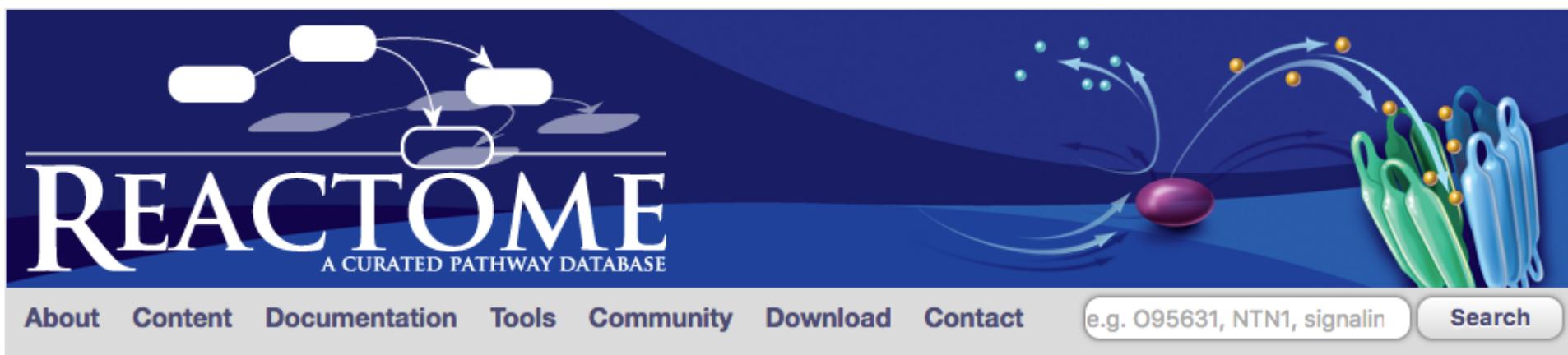


EMBL-EBI train online



Open Targets

# Data sources: Reactome

The image shows the Reactome homepage. At the top, there is a blue banner featuring a molecular interaction diagram with nodes and arrows. Below the banner, the word "REACTOME" is written in large, white, serif capital letters, with "A CURATED PATHWAY DATABASE" in smaller letters underneath. Below the title, there 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. The main content area below the banner contains a detailed molecular interaction diagram with various colored nodes (blue, green, yellow) and arrows indicating biological processes.

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

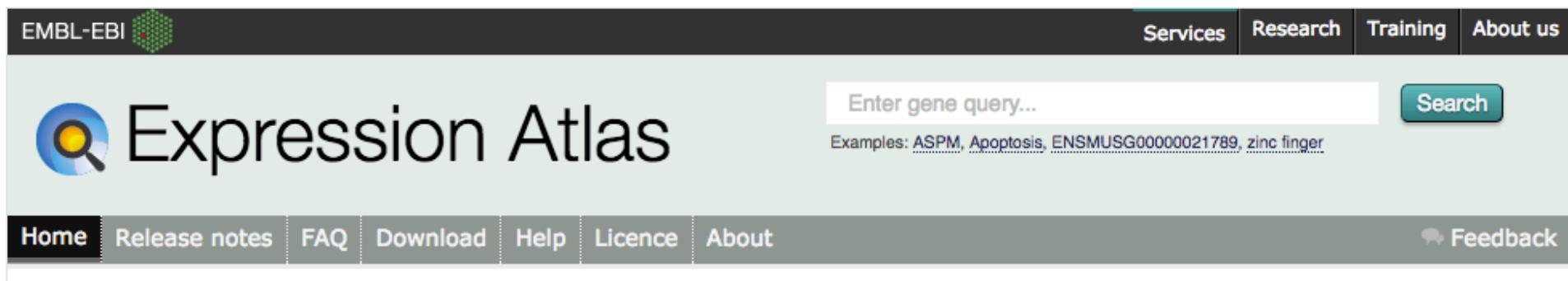


EMBL-EBI train online



Open Targets

# Data sources: Expression Atlas



The screenshot shows the Expression Atlas website. 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.

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

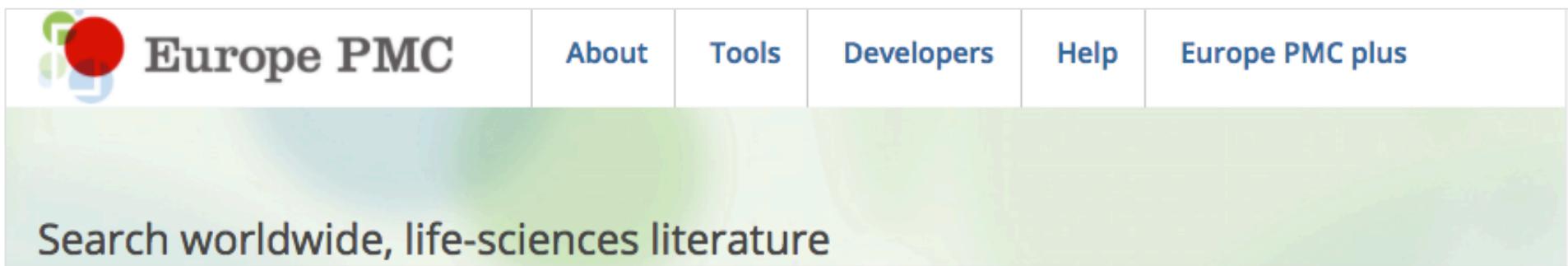


EMBL-EBI train online



Open Targets

# Data sources: Europe PMC



A screenshot of the Europe PMC website. The header features the Europe PMC logo (a stylized 'E' composed of blue and red circles) and the text "Europe PMC". Below the header is a navigation bar with links for "About", "Tools", "Developers", "Help", and "Europe PMC plus". A large green banner below the navigation bar contains the 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 is a dark header bar with the Wellcome Trust Sanger Institute logo on the left, featuring a stylized 'S' icon and the text 'wellcome trust sanger institute'. 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', 'Campus', and a magnifying glass icon for search. Below the header is a large, bold title: 'Welcome to PhenoDigm (PHENOtype comparisons for Disease and Gene Models)'. Underneath the title is a horizontal menu bar with three items: 'Diseases' (which is highlighted in blue), 'Tissue phenotype associations', and 'Secondary phenotypes'.

- Semantic approach to associate mouse models with diseases

# 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?

# 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](https://github.com/opentargets/association_score_methods)

# Factors affecting the relative strength of an evidence

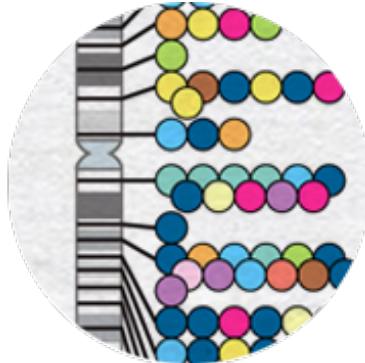
e.g. *GWAS Catalog*

$$S = f * s * c$$

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\*

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

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

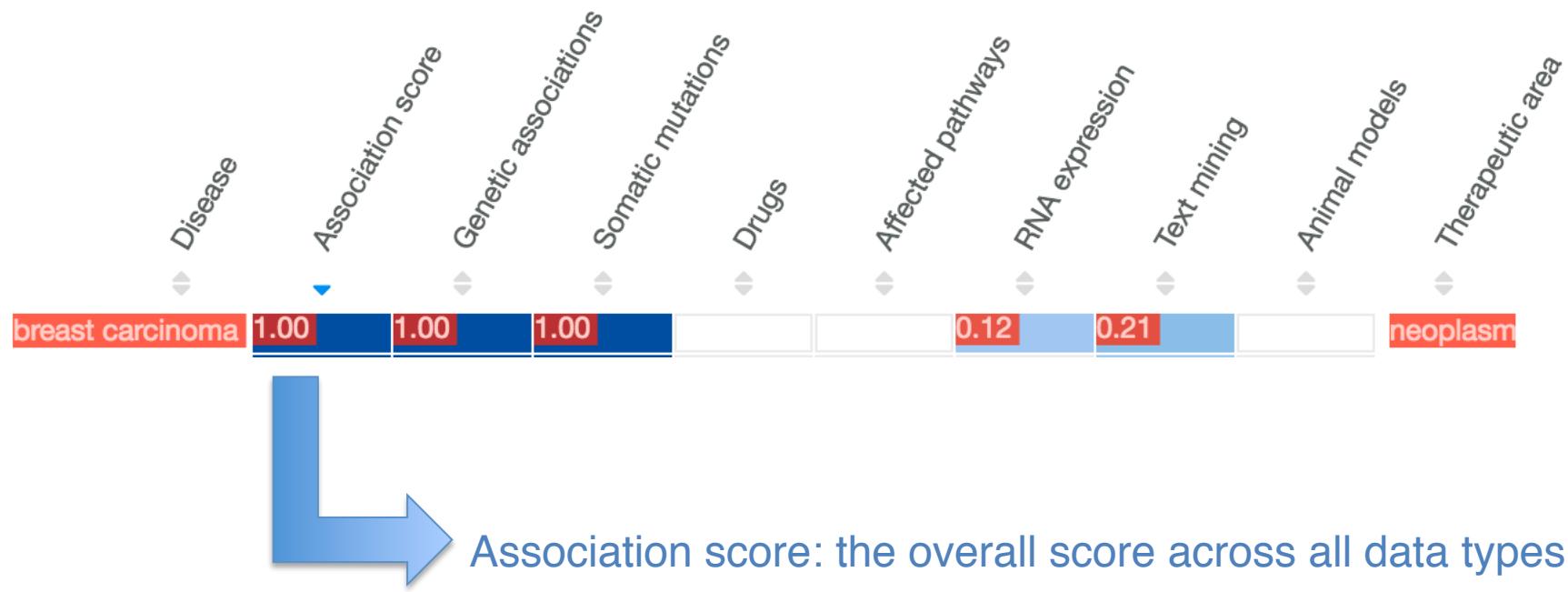
# Disclaimer: score, dos and don'ts

- 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

# 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?

# 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



Open/Hide | List operations | Expand operations

public : Publicly supported stable API.

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>

# Extra hands-on exercises

Pages 35-36

# 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

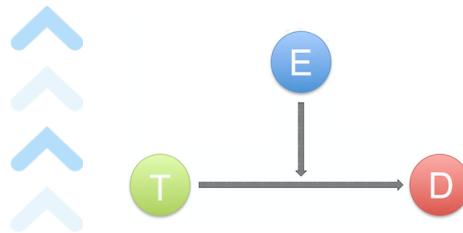
Oh Yes!  
And all is 100% free  
and open source



Open Targets

# Currently: Integration of existing data

## Public Databases and Pipelines



Open Targets experimental data: NEW  
Physiologically relevant and at scale

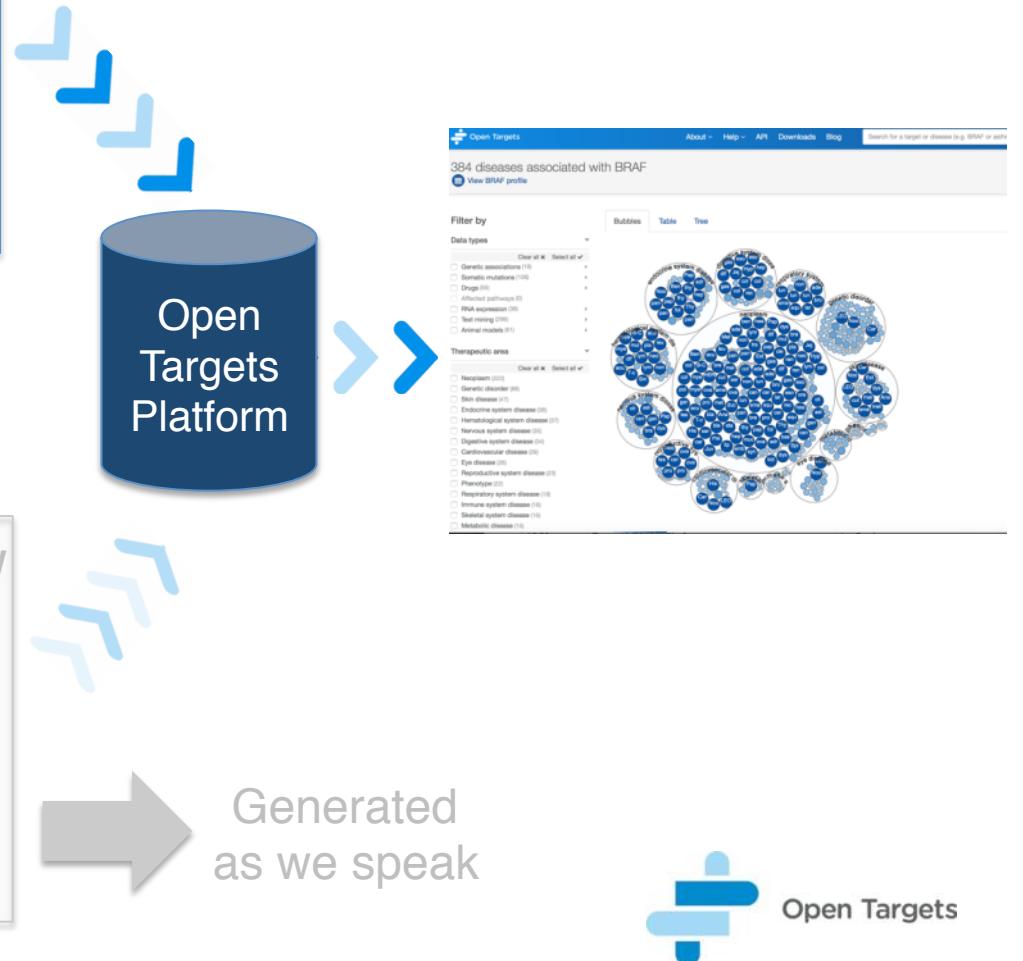
Oncology



Immunology



Neurodegeneration



Open Targets

# We support decision-making

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

# How to cite us

Published online 8 December 2016

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

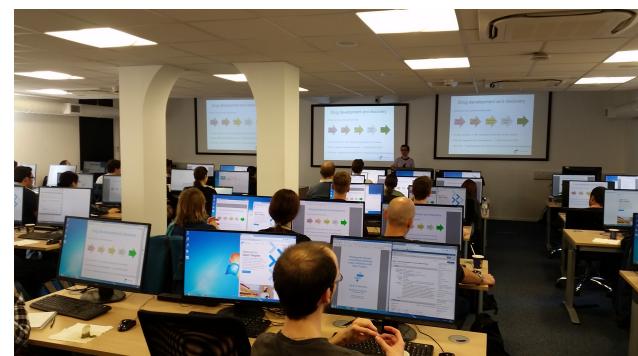
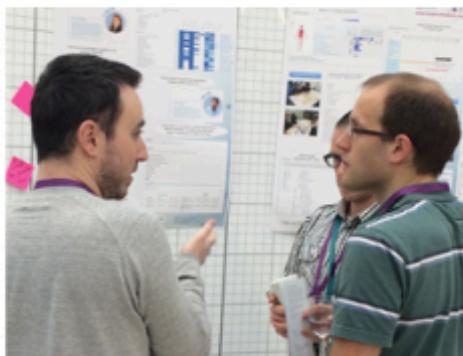
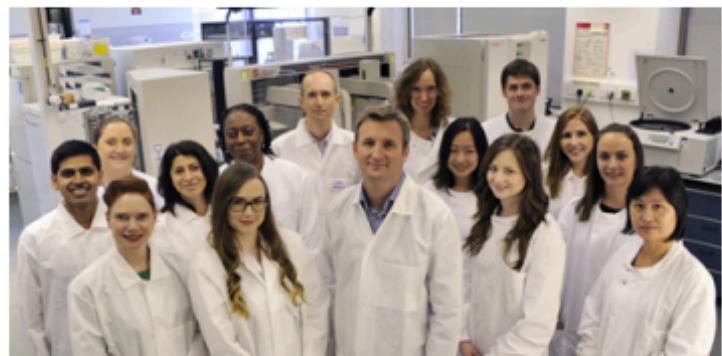
## Open Targets: a platform for therapeutic target identification and validation

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)



Open Targets

# Feedback

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

# Get in touch



@targetvalidate



support@targetvalidation.org



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



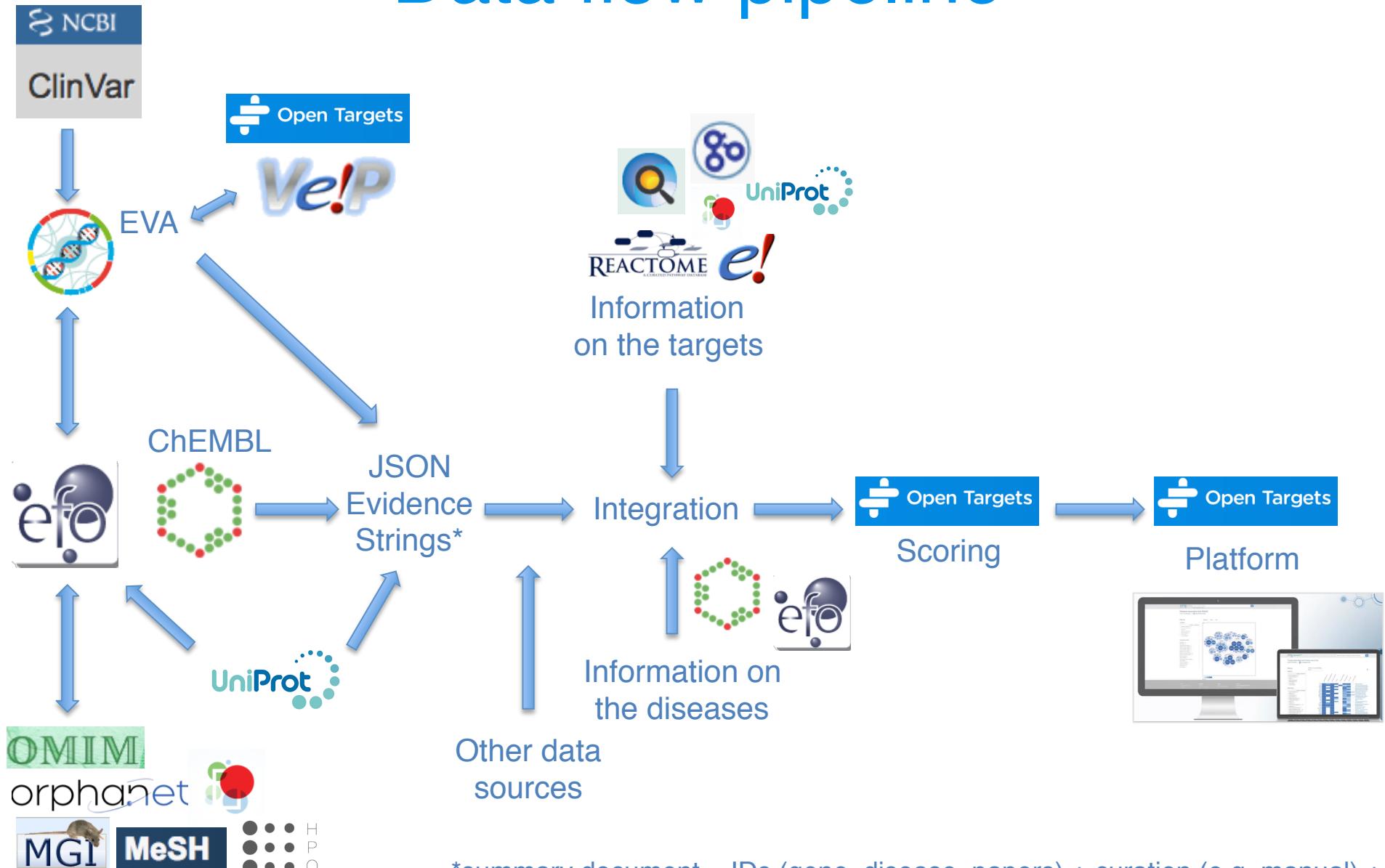
[blog.opentargets.org/](http://blog.opentargets.org/)



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



# Data flow pipeline



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