

Mining gene-disease associations and drug target validation with Open Targets

CRUK Cambridge Institute

Denise Carvalho-Silva

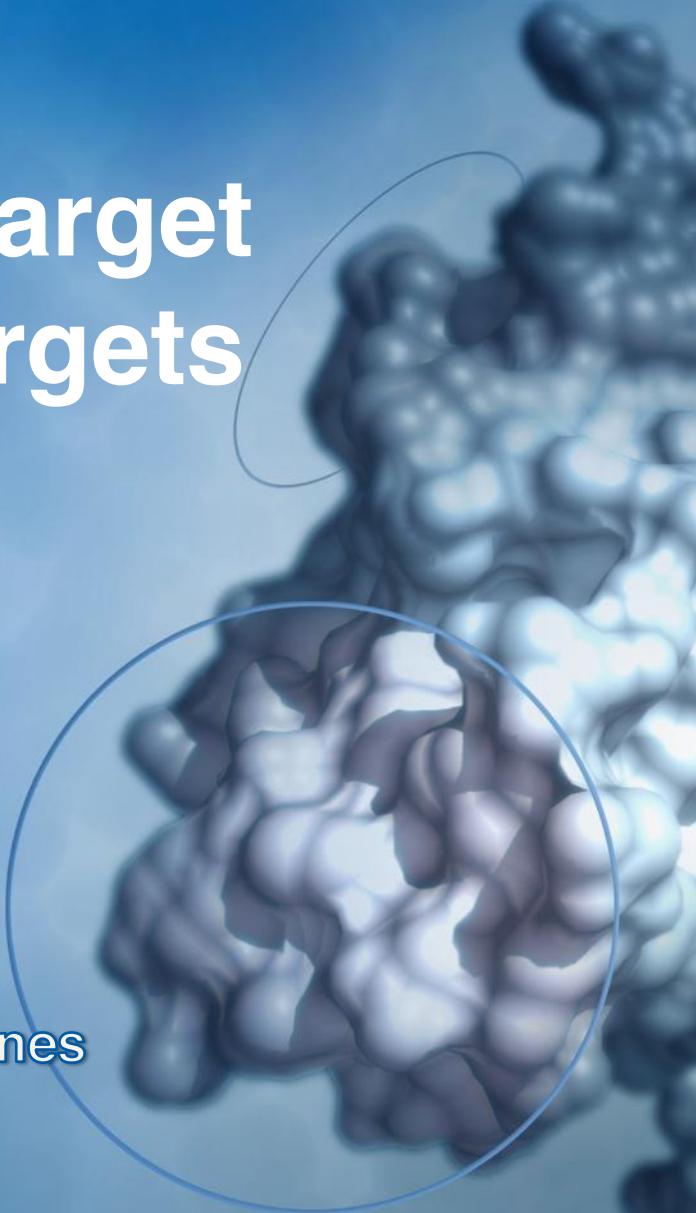
Wellcome Genome Campus

Open Targets

Core Bioinformatics and Computational Pipelines



Open Targets



Materials

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



slides

Open Targets



Hands-on Workshop
Course booklet

Open Targets



Hands-on Workshop
Answer booklet

exercises

answers



Today 13:00-16:00

- Introduction to the Open Targets Consortium
- The Open Targets Platform

Live demos, talks, hands-on exercises

- Wrap up and feedback survey

Course objectives

What is the Open Targets Consortium?

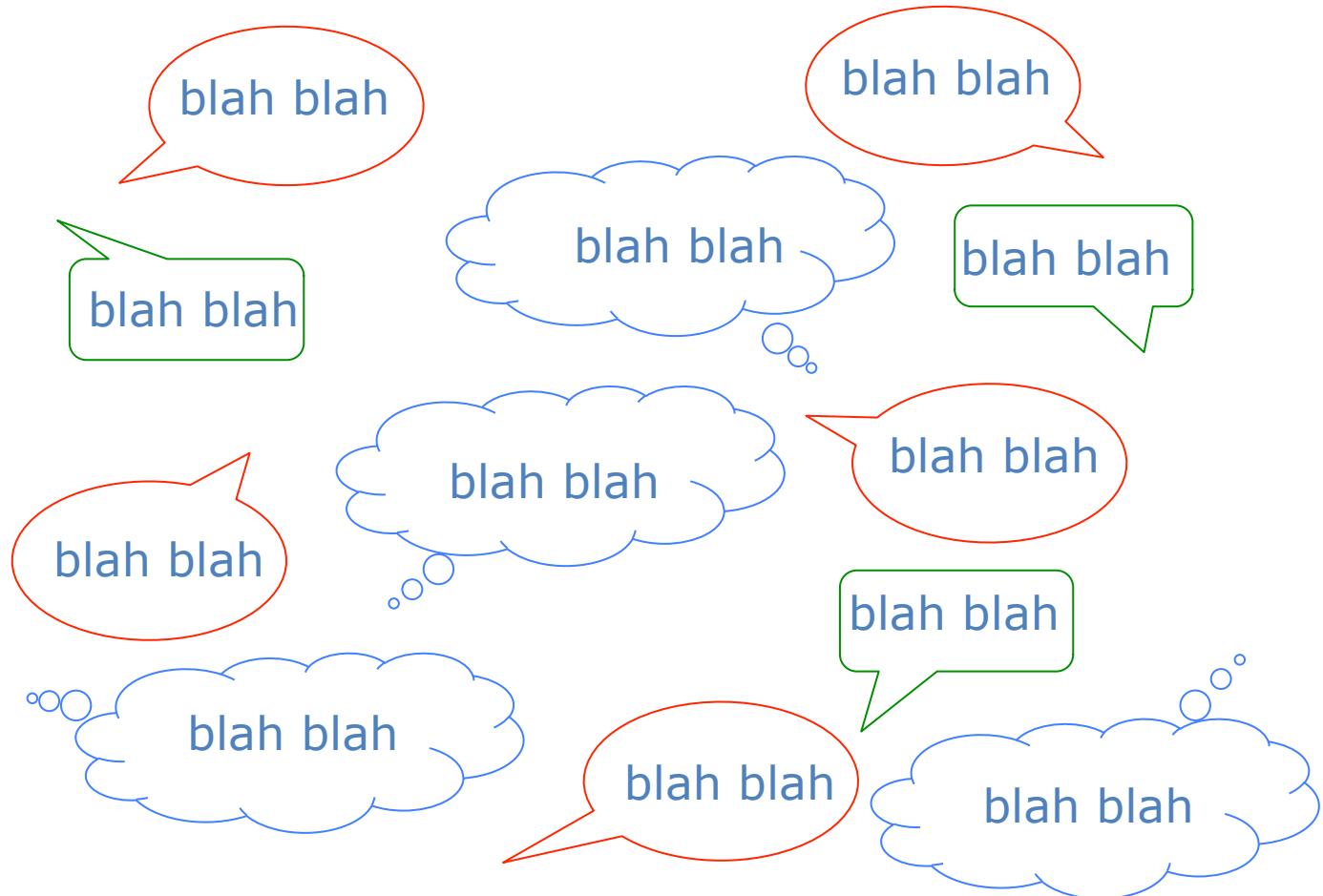
What is the Open Targets Platform?

How to navigate the Platform?

How to connect with us



Open Targets



Open Targets

Do we need programmes for drug development and discovery?

YES

NO

Quora Ask or Search Quora Ask Question

Infectious Diseases Medical Conditions and Diseases +1

How many diseases today have **no** cure? Which ones?



Karen Tiede, Hula hoop for your health.

641 Views · Most Viewed Writer in Infectious Diseases with 90+ answers

How many? **Most of them, probably.** We have treatments, and support, and care, and some of the time, you live through the disease, but it's not exactly that we "cured" it. You simply didn't have to die of it.

To add to Meghana's list,* diabetes (both kinds), heart disease, arthritis, pretty much all of the auto-immune disorders, most of the mental illnesses, many of the genetic disorders (cystic fibrosis, for example).

The list of diseases that can actually be cured, rather than prevented or treated, is pretty short.

Written Aug 9, 2014 · View Upvotes · Answer requested by 1 person

Upvote | 4

Downvote Comment

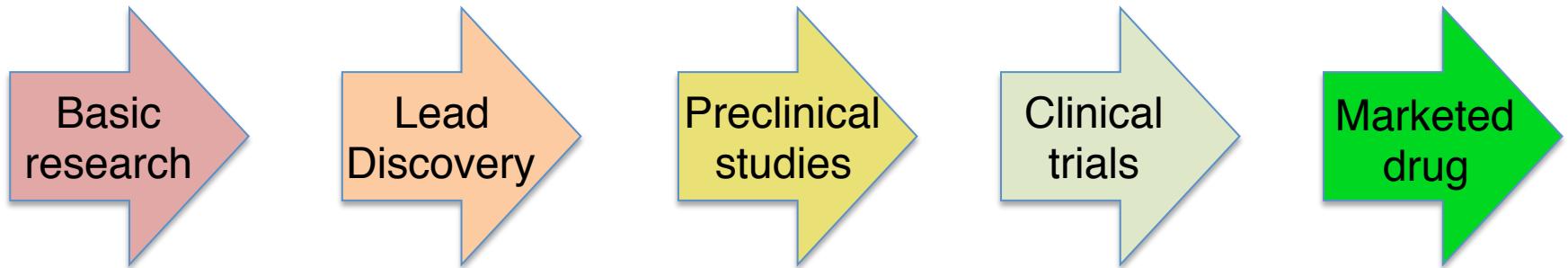


* **Meghana Rastogi**, Research Scholar

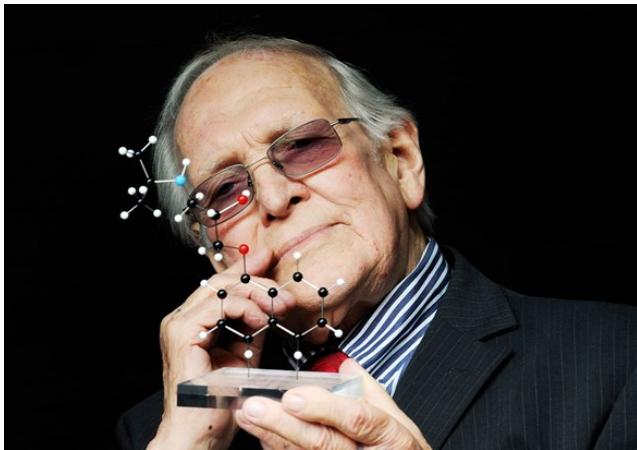
Ebola, HIV, Cancer, Dystonia are one of those disease which don't have any cure.

Drug development and discovery

- Driven by this unmet clinical need



- 10 year journey → new drug from discovery to the market
- Cost for research and development: ~ 2,000,000,000 GBP
- 10% only → approved by the FDA (U.S Food and Drug Administration)



Sir James Black, winner of the 1988 Nobel Prize in Physiology and Medicine for his work in drug development e.g. **propranolol** (beta blocker) and **cimetidine** (histamine H₂ receptor antagonist).

“The most fruitful basis for the discovery of a **new** drug is to start with an **old drug**”.

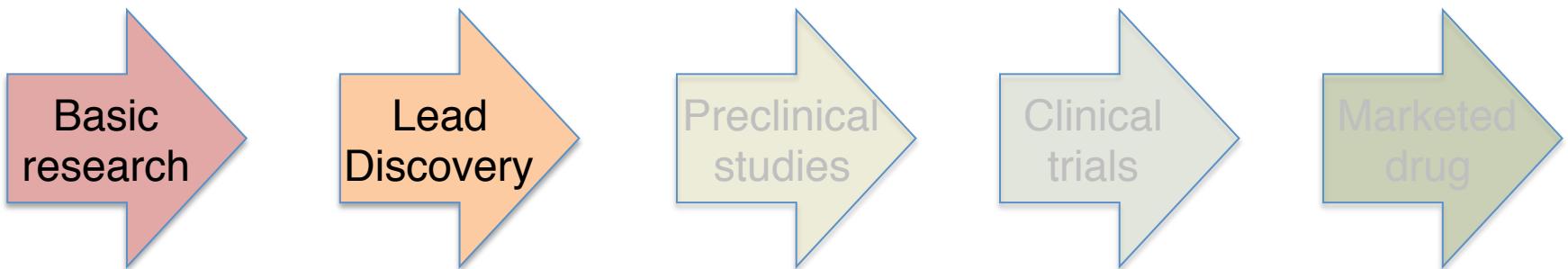
Drug repurposing

Drugs do fail in the clinic

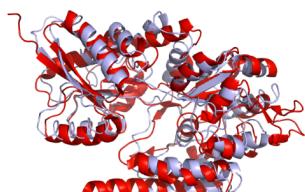
They don't work

They are not safe

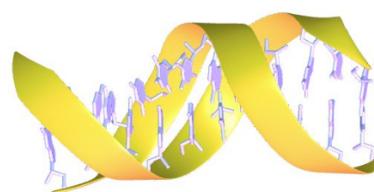
Important role of target identification and validation



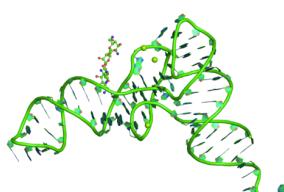
But what is a target?



protein



DNA



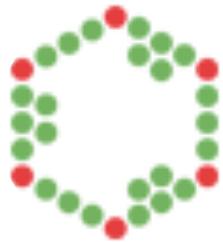
RNA



Open Targets

Public databases for drug discovery

- European Bioinformatics Institute (EBI-EMBL)



- Supporting all stages of drug discovery
- ‘Fitting together like a jigsaw puzzle’ to build hypotheses



Open Targets

Huge hurdle: data is everywhere



Wouldn't it be great to have all these data **integrated** in one place?



Yes, one **one-stop shop** i.e. a database with **comprehensive** and **trustworthy** data that we all could access it for **free**?



That'd be fab! It'd be much quicker to carry out our experiments in the lab **identifying** and **prioritising** new drugs.



Open Targets



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



*Patrick Vallance, President
Pharmaceuticals R&D
GlaxoSmithKline*



*Professor Sir
Mike Stratton
Director, Sanger Institute*



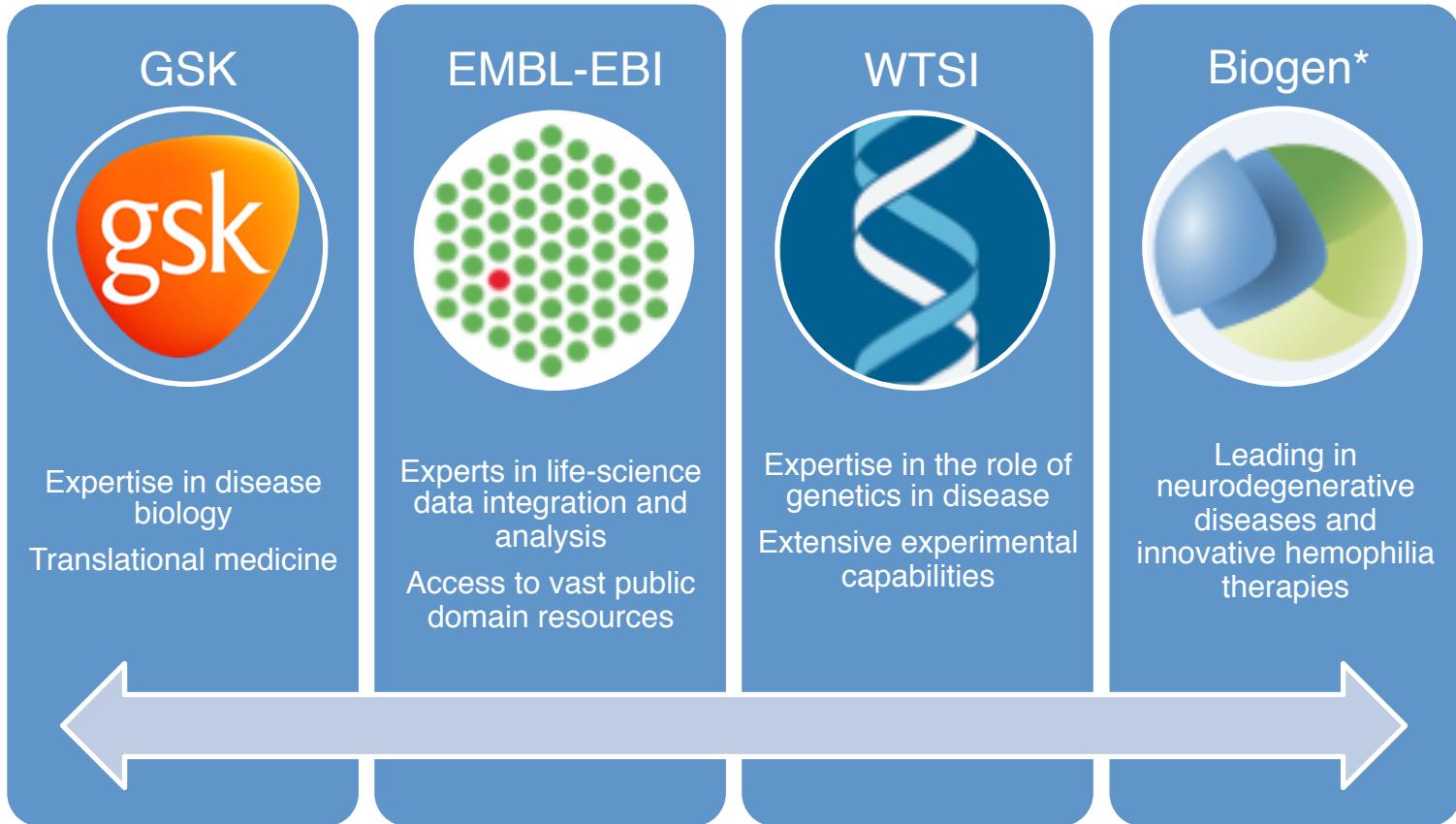
- Target validation can (should) be improved
- One institution could not necessarily or easily do it alone
- Strong desire to collaborate based on highly complimentary skills set
- Existing strong relationships, real commitment to the mission

The consortium

- Launched in March 2014
- Formerly known as CTTV
- Rebranded to Open Targets in April 2016
- Public-private initiative, precompetitive, rapid publication
- Aim: **transform** drug discovery
- How? Through the **identification** and **prioritisation** of targets

<http://www.ebi.ac.uk/about/news/press-releases/open-targets-new-name-new-data>

Who is Open Targets?



*Biogen joined the consortium in February 2016

The two major areas of work within Open Targets

Core bioinformatics pipelines Part I*



Database for data integration
Web portal
REST API
Python client (fully supported)
R client (community)
Data dumps

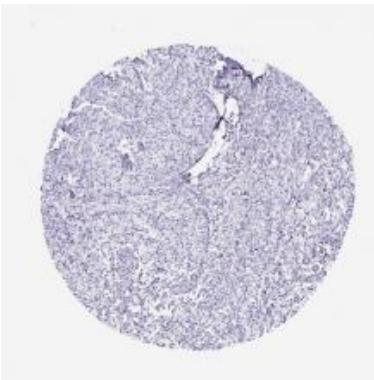
Experimental projects Part II



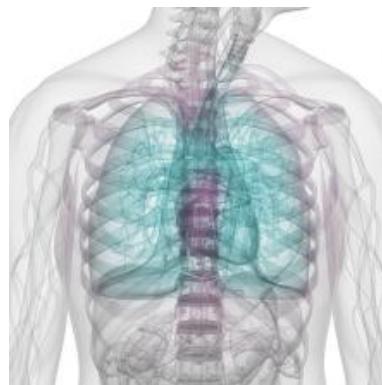
Generate new data
CRISPR
Organoids
Single cell RNASeq
Cell line fusion analyses
Metabolite GWAS

www.opentargets.org/projects

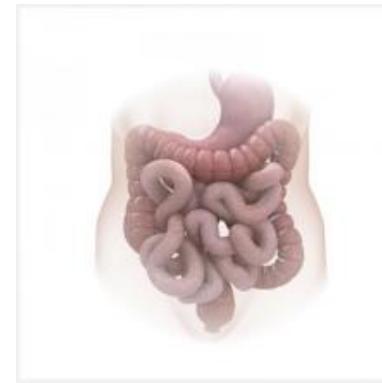
Experimental projects



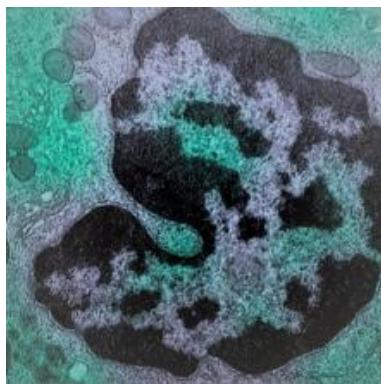
Oncology



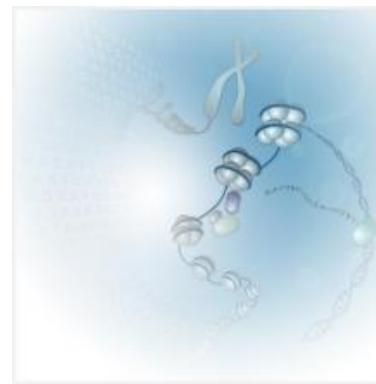
Respiratory



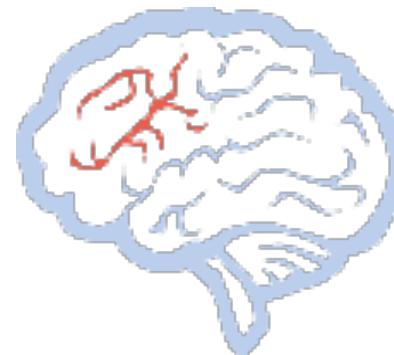
IBD



Inflammation
and immunity



Cell line
epigenomes



Neurodegenerative
diseases

Scientific project portfolio

Oncology



Immunology



Neurodegeneration



Cross-Disease



Human cellular experiments

- 1000 Cell-line Fusion Analysis
- CRISPR synthetic lethality screen
- NK cell receptors

- IBD Organoids,
- MacroScreen
- Dendritic Cell Screen
- Asthma Single-cell

- Ox. stress & tau CRISPR
- Familial Alzheimer's single-cell RNAseq



Genetics as a tool

- NGS Melanoma
- Cancer Signaling Pathways

- IBD & MS GWAS fine-mapping
- Bronchiectasis

- Parkinson's & Alzheimer's GWAS resolution

- Influential Variants
- Metabolite GWAS



Enabling resource

- Immune cell functional maps
- IBD BioResource

- Neuron functional maps

- CELLector
- Cell line Epigenomics



Robust Data Integration
www.targetvalidation.org

Browsing the Open Targets website

opentargets.org

Can you explore Open Targets consortium website to find out:

- More about the consortium, including its core principles
- The Open Targets projects
- Types of cancer under experimental investigation
- The key challenge of the Core Bioinformatics team
- How to get to the the Open Targets Platform

Core bioinformatics pipelines

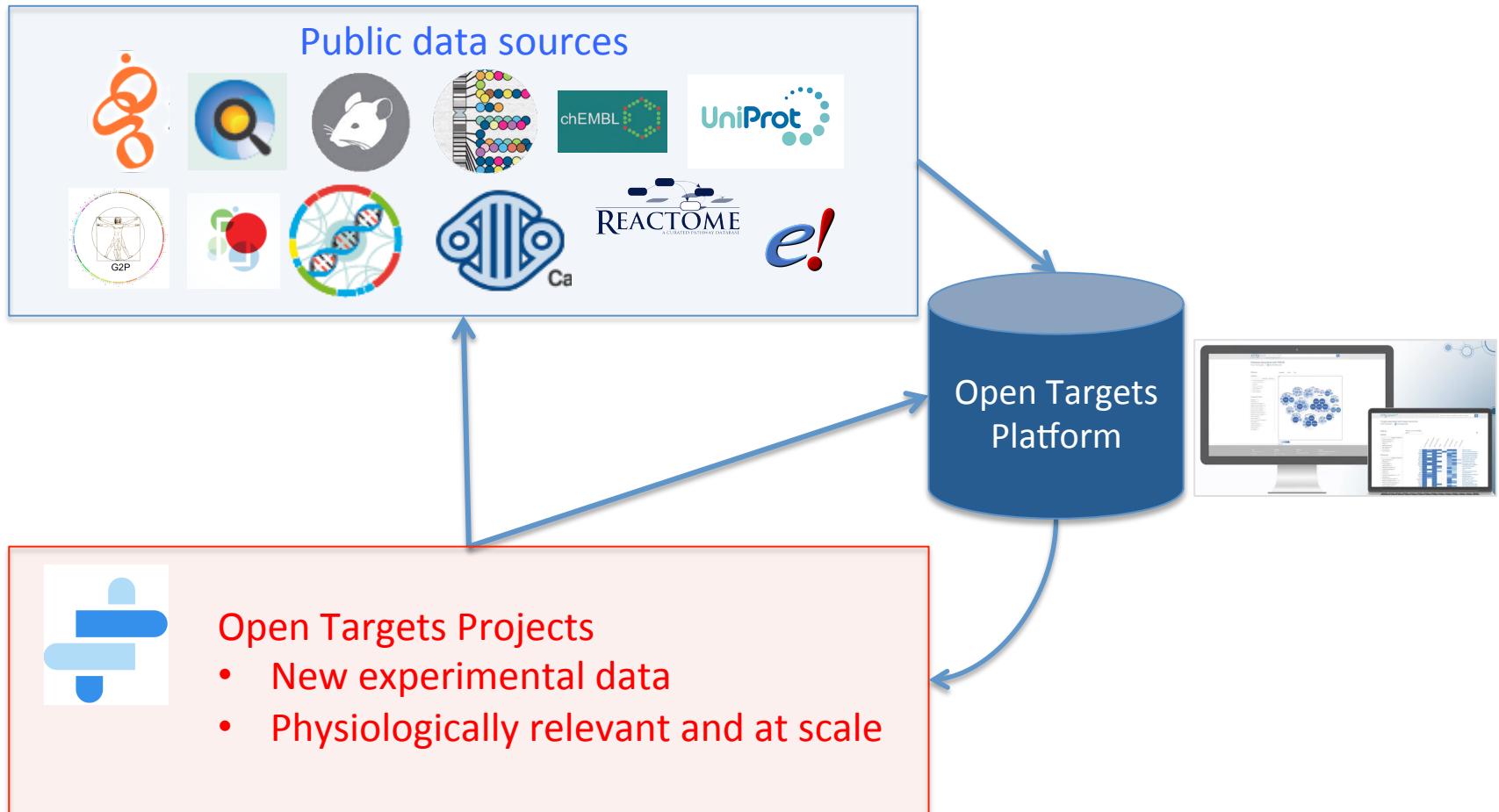
- Open Targets Platform: first release (Dec 2015)

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



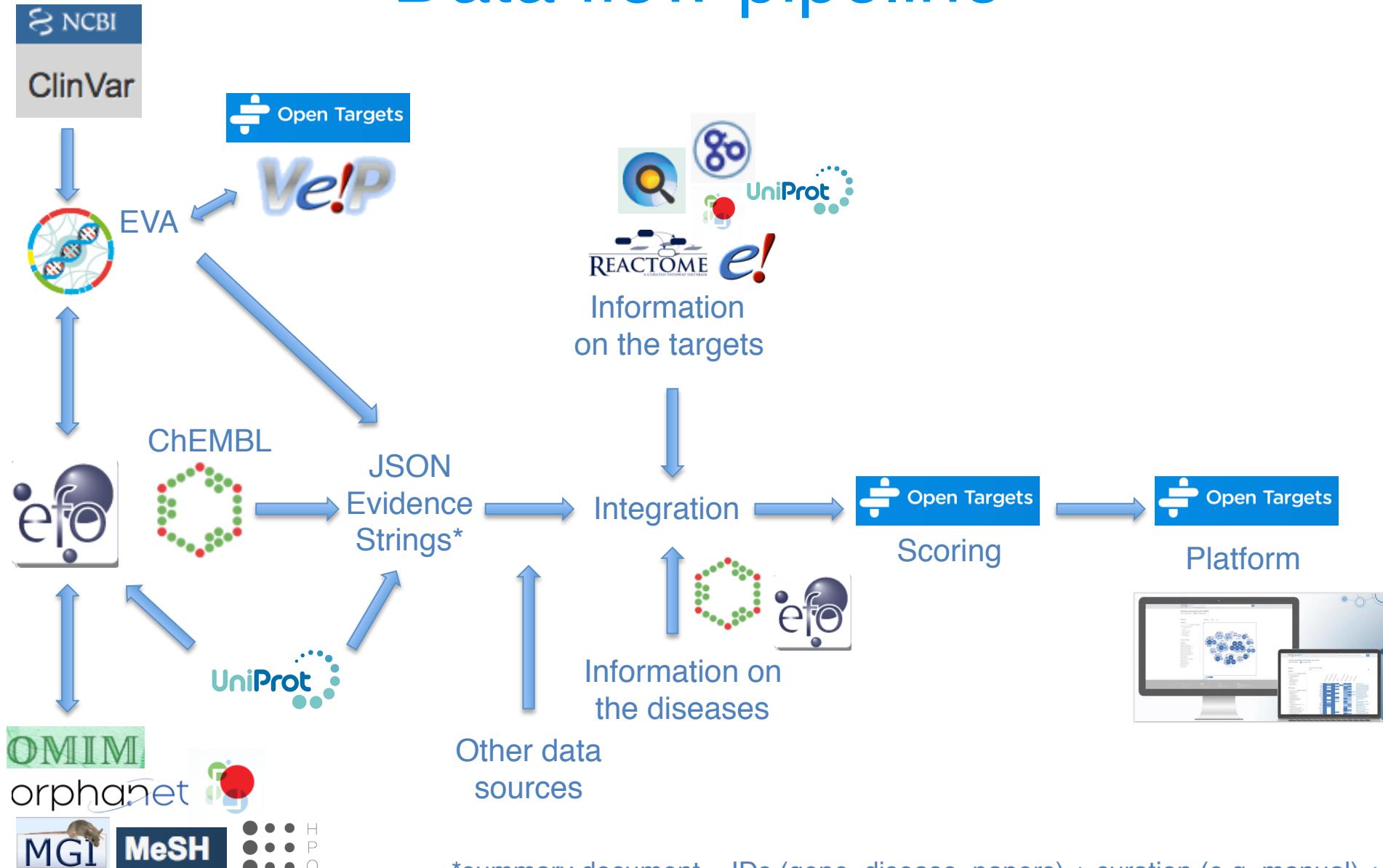
- Help scientists identify and prioritise relationships between targets and diseases

Platform to integrate existing and new data



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

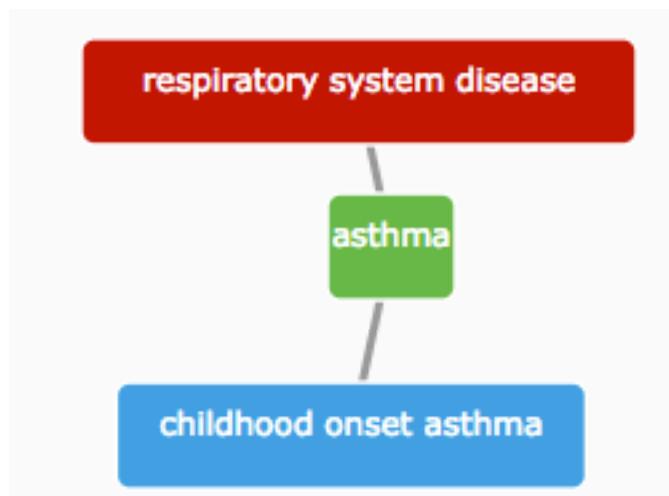
Data flow pipeline



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

Experimental Factor Ontology (EFO)

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



controlled vocabulary
+
hierarchy (relationship)

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

How do we associate diseases w/ targets?

Step 1: mine phenotypes and diseases from ChEMBL, UniProt and EVA (w/ ClinVar)

Step 2: map these to an ontology using EFO terms

Step 3: use genes as proxies for our targets

Step 4: create target-disease evidence JSON objects

Step 5: calculate for each evidence the likelihood of gene A being associated with disease B

Step 6: compute integrated target-disease scores at the data source, data type and overall levels

Hypothetical scenario for a use case

The screenshot shows the header of the *Nature Genetics* website. The logo "nature genetics" is on the left. To the right is a decorative background image of a DNA double helix. Below the logo is a horizontal navigation bar with links: Home, Current issue, Comment, Research, Archive (with a dropdown arrow), Authors & referees (with a dropdown arrow), and About the journal (with a dropdown arrow).

[home](#) ▶ [archive](#) ▶ [issue](#) ▶ [letter](#) ▶ [full text](#)

NATURE GENETICS | LETTER



Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing

Alejandro Sifrim, Marc-Phillip Hitz, Anna Wilsdon, Jeroen Breckpot, Saeed H Al Turki, Bernard Thienpont, Jeremy McRae, Tomas W Fitzgerald, Tarjinder Singh, Ganesh Jawahar Swaminathan, Elena Prigmore, Diana Rajan, Hashim Abdul-Khalil, Siddharth Banka, Ulrike M M Bauer, Jamie Bentham, Felix Berger, Shoumo Bhattacharya, Frances Bu'Lock, Natalie Canham, Irina-Gabriela Colgiu, Catherine Cosgrove, Helen Cox, Ingo Daehnert, Allan Daly
+ et al.

Nature Genetics 48, 1060–1065 (2016) | doi:10.1038/ng.3627

Received 23 December 2015 | Accepted 24 June 2016 | Published online 01 August 2016

Congenital heart disease (CHD)

targetvalidation.org

- How many targets are associated with CHD?
- Can you filter this number to get the targets based on Genetic associations only?
- Which data sources were used to support this genetic association?

Data sources and types

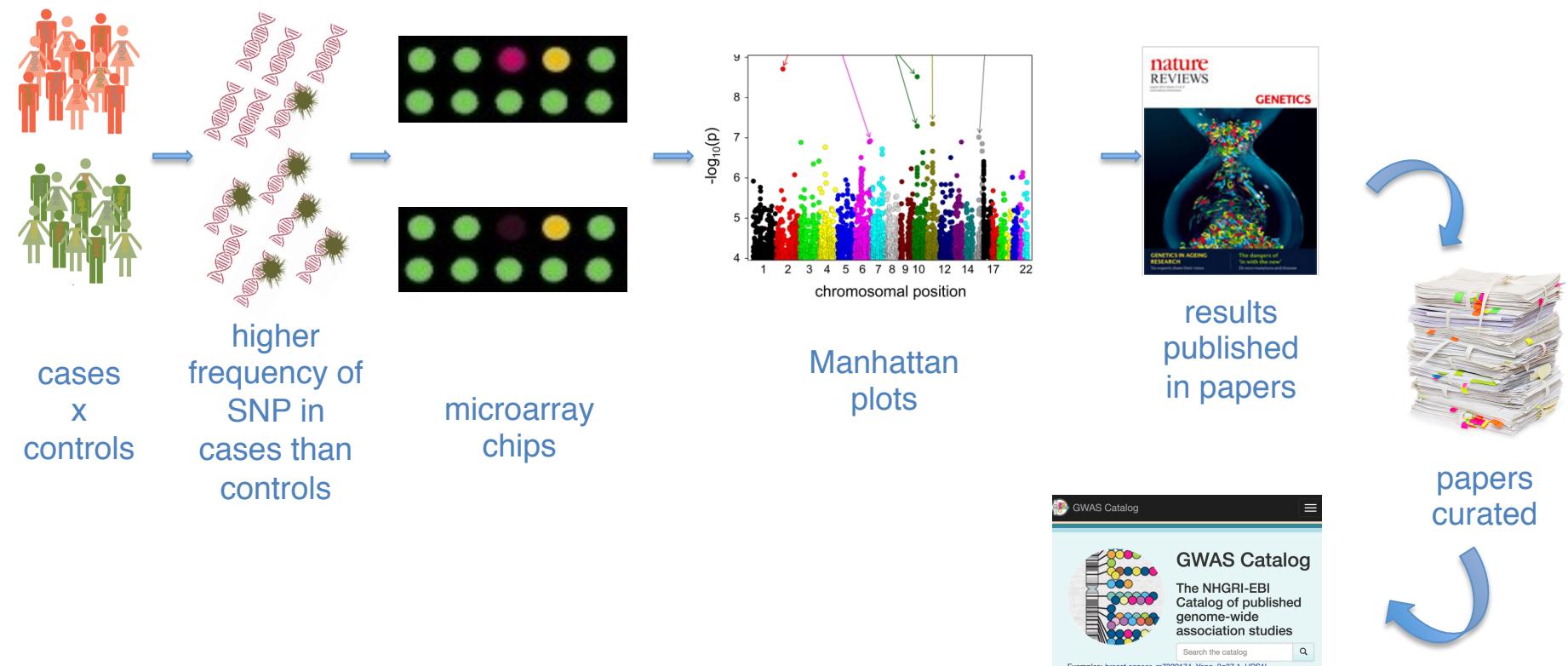
- Similar data sources are grouped for the calculation of the score

Data sources	Data type
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

- Do you have a favourite data you would like us to have?

Data sources: GWAS catalog

- GenomeWide Association Studies (www.nature.com/nrg/series/gwas/index.html)
- Array-based chips → genotyping 100,000 SNPs genomewide



<https://www.ebi.ac.uk/gwas/>

SNP-trait associations
Published in the GWAS catalog

Data sources: UniProt* and UniProt literature**

- Catalog of protein information (sequence, annotation, function)



* Manual curation of variants in the coding region, seen in patients

** Associations between target and disease, no specific mutation

Data sources: EVA

- Catalog of genetic variants (SNPs, CNVs; germline or somatic)
- Clinical information from ClinVar available: rare diseases

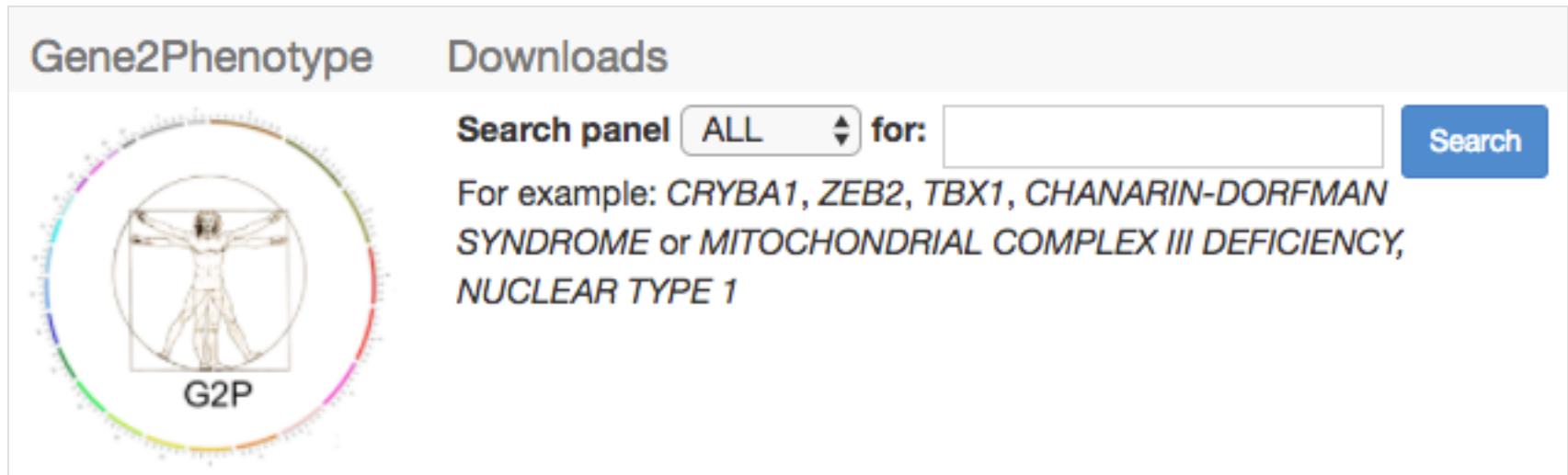
The screenshot shows the European Variation Archive (EVA) website. The header features the EVA logo and the text "European Variation Archive". Below the header is a navigation bar with links: Home, Submit Data, Study Browser, Variant Browser, Clinical Browser (which is highlighted in blue), GA4GH, API, FAQ, and Feedback.

The main content area is titled "ClinVar Browser" with a help icon. On the left, there is a "Filter" sidebar. The "Position" filter is set to "Assembly : GRCh37". The "Filter By:" dropdown is set to "Chromosomal", and the "Chromosome" dropdown shows "2:48000000-49000000". The "Consequence" filter has a "search" input field.

The main table displays 10 records out of 960, showing clinical variant details:

...	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	C/T	5_prime_UTR...	Lynch synd...	Uncertain s...	RCV000...
2	480...	MSH6	C/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	Benign	RCV000...
2	480...	MSH6	G/T	5_prime_UTR...	Hereditary ...	conflicting ...	RCV000...

Data sources: Gene2Phenotype



Gene2Phenotype

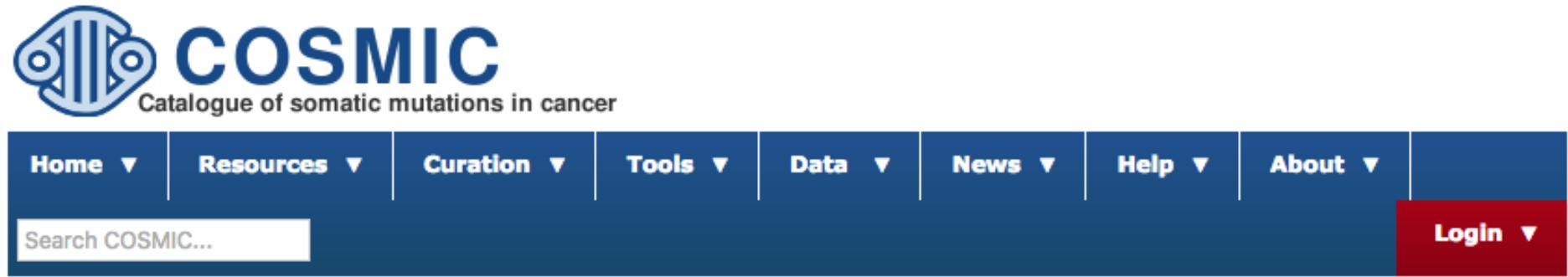
Downloads

Search panel ALL for: Search

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

- Catalog of variants, genes, phenotypes
- Developmental disorders
- Literature curation → consultant clinical geneticists in the UK

Data sources: The Cancer Gene Census



The screenshot shows the COSMIC website homepage. At the top left is the COSMIC logo, which consists of a stylized blue eye-like icon followed by the word "COSMIC" in a bold, blue, sans-serif font. Below the logo is the tagline "Catalogue of somatic mutations in cancer". A horizontal navigation bar follows, featuring eight items: "Home ▾", "Resources ▾", "Curation ▾", "Tools ▾", "Data ▾", "News ▾", "Help ▾", and "About ▾". To the right of this bar is a red "Login ▾" button. Below the navigation bar is a search bar containing the placeholder text "Search COSMIC...".

- catalog of genes for which mutations have been causally implicated in cancer
- genes associated with specific plus other cancers associated with that gene

Data sources: IntOGen

The screenshot shows 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 main header area featuring the IntOGen logo (an orange stylized 'i' icon) and the text "IntOGen Integrative Onco Genomics". A search bar contains the placeholder text "e.g. Mutation frequency of VHL" and includes a microphone icon for voice search. Below the search bar are links for "Search example" and "Show more examples".

intOGen

Search Downloads Analysis About

Sign In

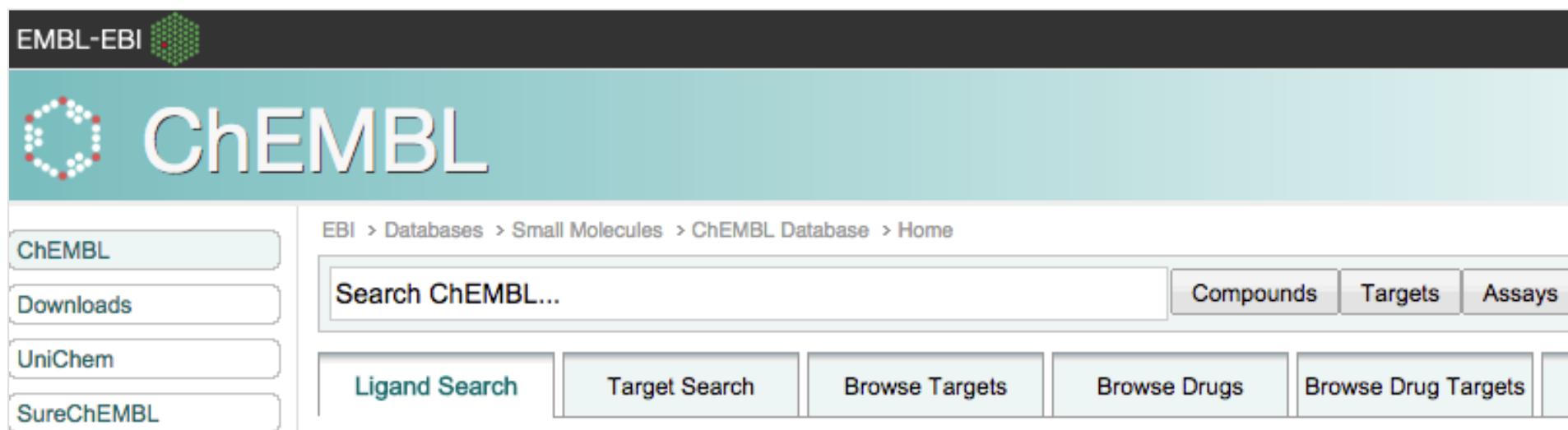
Integrative
Onco
Genomics

e.g. Mutation frequency of VHL

Search example | Show more examples

- catalog of genes and somatic (driver) mutations
- involvement in cancer biology

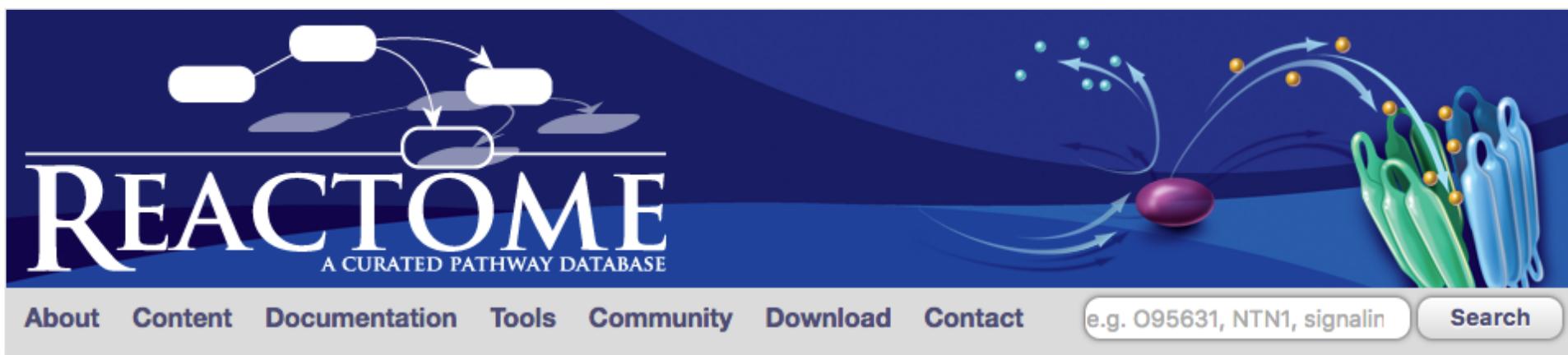
Data sources: ChEMBL



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 pattern) 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...". Above the search bar is a breadcrumb trail: "EBI > Databases > Small Molecules > ChEMBL Database > Home". To the right of the search bar 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".

- Catalog of known drugs linked to a disease and a known target
- Drugs → FDA approved for marketing or clinical trials

Data sources: Reactome

The image shows the Reactome homepage. At the top, there is a large blue banner featuring a molecular interaction diagram with nodes and arrows. Below the banner, the word "REACTOME" is written in large white letters, with "A CURATED PATHWAY DATABASE" in smaller text 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.

- Catalog of biochemical reactions and pathways
- Manual curation of pathways affected by mutations

Data sources: Expression Atlas

EMBL-EBI 

Services Research Training About us

 Expression Atlas

Enter gene query...
Examples: [ASPM](#), [Apoptosis](#), [ENSMUSG00000021789](#), [zinc finger](#)

Search

Home Release notes FAQ Download Help Licence About Feedback

Expression Atlas: Differential and Baseline Expression

The Expression Atlas provides information on gene expression patterns under different biological conditions. Gene expression data is re-analysed in-house to detect genes showing interesting baseline and differential expression patterns. [Read more about Expression Atlas.](#)

- Catalog of patterns of gene expression
- Baseline expression for human genes
- Differential expression (healthy versus diseases tissues)

Data sources: Europe PMC

The screenshot shows the Europe PMC homepage. 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 overlapping colored circles (blue, green, and red) followed by the text "Europe PMC". Below the navigation bar is a search bar with the placeholder text "Search worldwide, life-sciences literature". To the right of the search bar is a blue "Search" button with a magnifying glass icon. Below the search bar, there is an example search query: "E.g. "breast cancer" HER2 Smith J".

- Text mining → association between targets and diseases
- Titles, abstracts, full text (but supplementary tables) are mined
- Co-occurrence in the same sentence of target and disease names (or synonyms)

Data sources: PhenoDigm

The screenshot shows the homepage of the PhenoDigm database. 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 containing links for "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 a search function. 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 and underlined), "Tissue phenotype associations", and "Secondary phenotypes". Below this menu, a large text block reads: "Analyzing curated phenotype annotations to associate animal models with human diseases".

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

Diseases Tissue phenotype associations Secondary phenotypes

Analyzing curated phenotype annotations to associate animal models with human diseases

- Semantic approach to map between clinical features observed in humans and annotations of phenotypes in mouse models
- PMID: 23660285

<http://www.sanger.ac.uk/resources/databases/phenodigm/>

Congenital heart disease (CHD)

- Let's focus on the top gene in the list, i.e. *GDF1*
- Can you get a list of genetic variants (i.e. mutations) that associate this target with CHD?
- How many papers support the association through text mining?
- Let's now focus on the target itself. Which tissues does this gene seem to be highly expressed according to the GTEx project?
- Are there other cardiovascular diseases associated with this target? How strong is this association?

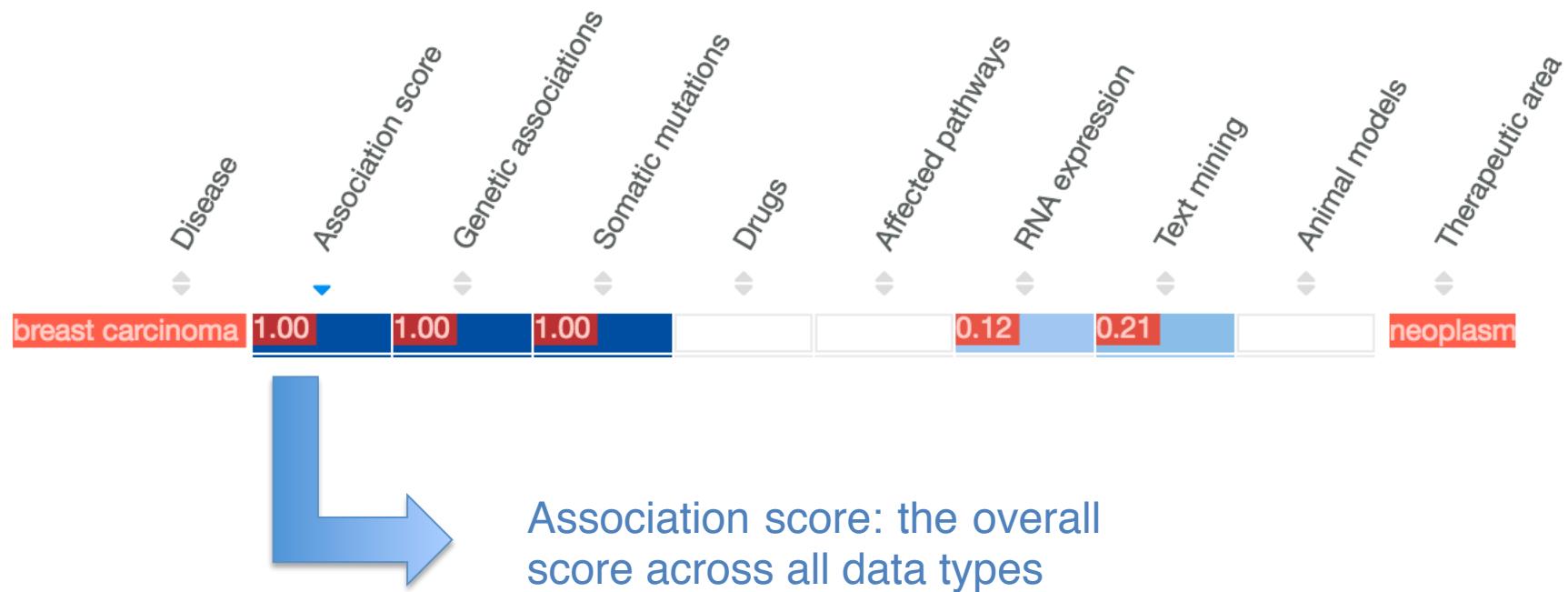
How confident can you be of the target-disease associations in Open Targets?

We've developed a scoring scheme*

- 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

*confidence, frequency, severity, aggregation
https://github.com/CTTV/association_score_methods

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 strong are the associations?

- Look for our score and the 5 shades of blue!



PIK3CD	1.00	1.00						0.03	0.18		
PIK3R1											

- No evidence to support the association: score is 0
- Sources providing their own score:  

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 hypotheses
- It can help you to decide on which target to pursue
- It is NOT sufficient on its own (use it in combination with...)

Drugs marketed and in clinical trials

One of the targets in the paper that is associated with CHD is *PRKD1*.

- What are the drugs currently in clinical trials and targeting this gene?
- Can you list a couple of cancer types where this gene has been targeted at?

Exercises and Answers

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



Wrap up

Open Targets Platform is the place:

Identify and prioritise targets for drug discovery

Target-disease associations: different sources

Integrated information on target and diseases

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



Open Targets

Open Targets Platform

- Intuitive and easy-to-use web interface
- Release 3-4 months: new data, new web features*



- Improvements driven by our user communities

* <https://www.targetvalidation.org/release-notes>

We support decision-making

A) Which targets are associated with a disease?

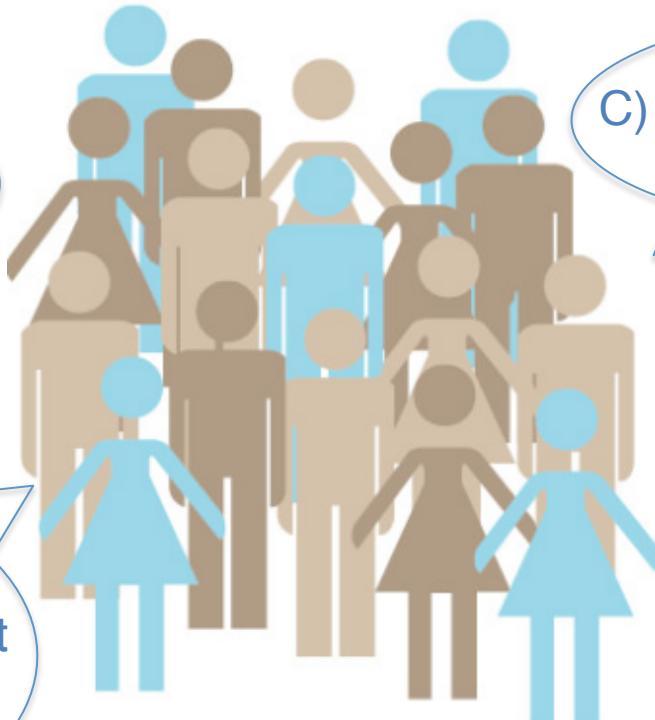
B) What evidence supports this target-disease association?

F) What else can I find out about my drug target?

E) If this target is associated with other diseases, can I get the association for diseases from different therapeutic areas?

C) Are there FDA drugs for this association?

D) For a given target, are there other diseases associated with it?



Summary statistics

- Latest release: December 2016 *

31,071
targets

2,559,080
associations

8,659
diseases

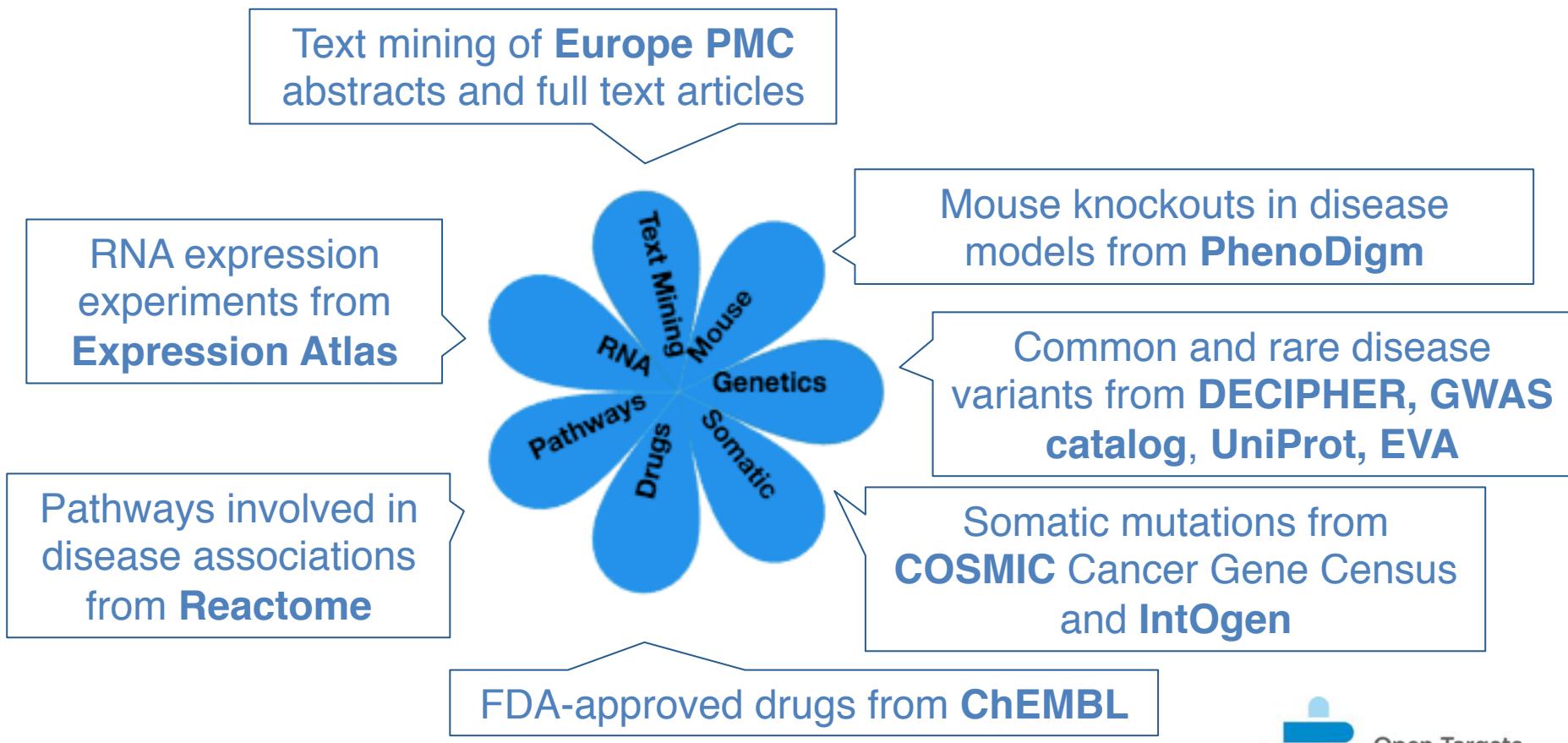
13
data sources



* <https://blog.opentargets.org/open-targets-platform-our-new-release-is-out-2/>

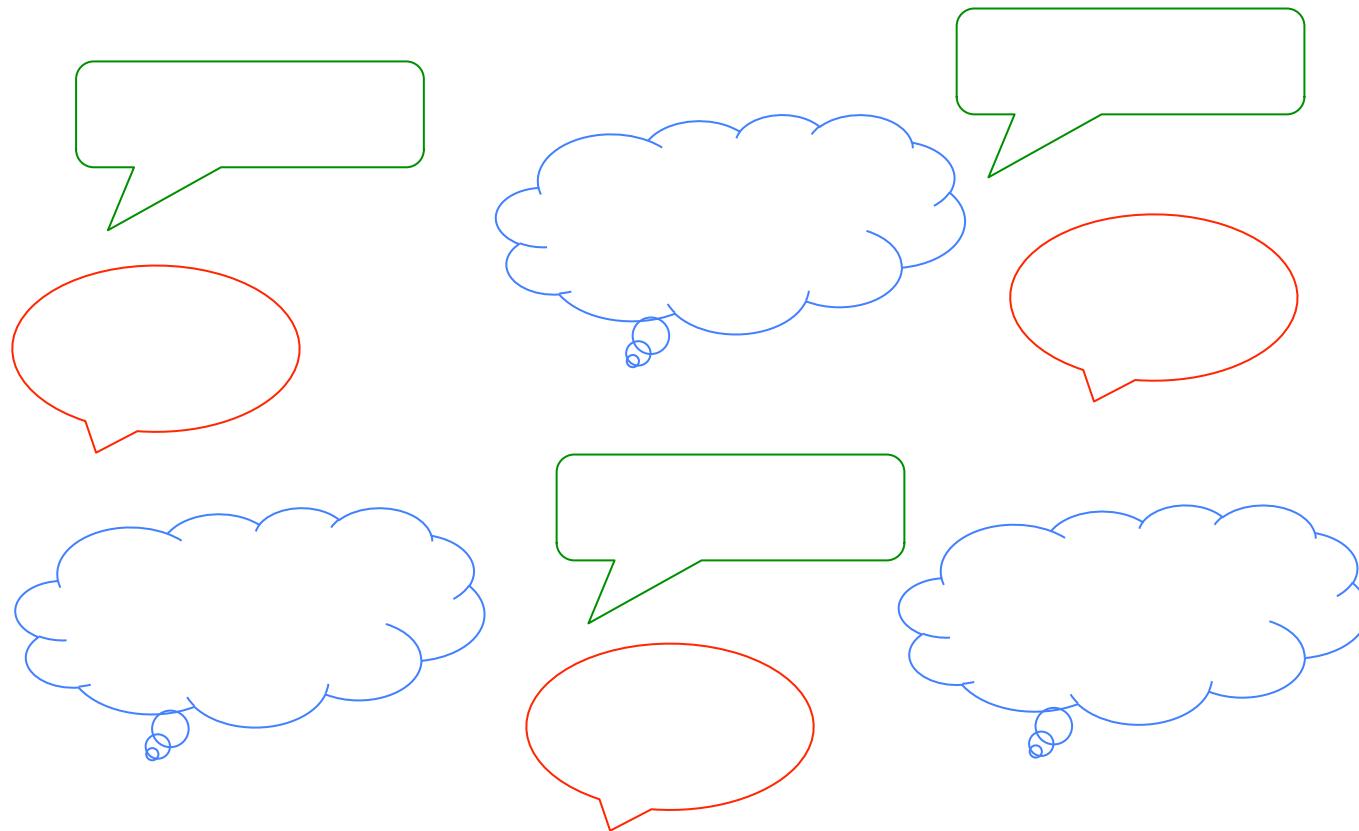
> 2 million associations

Supported by publicly available data: the evidence



Open Targets

Your take home message



Open Targets

Short feedback survey

<http://tinyurl.com/cam-161216>

Help and documentation

Get in touch



@targetvalidate



support@targetvalidation.org



www.facebook.com/OpenTargets/

The Target Validation Platform:

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

Frequently Asked Questions:

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

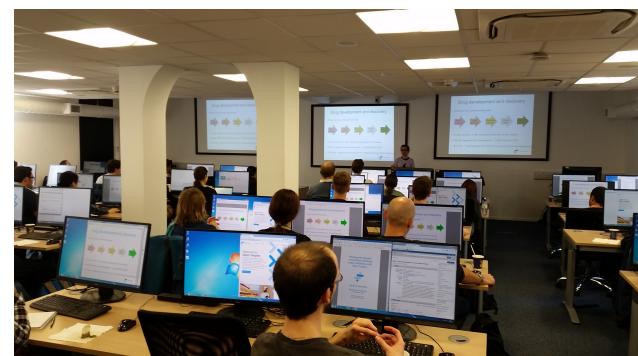
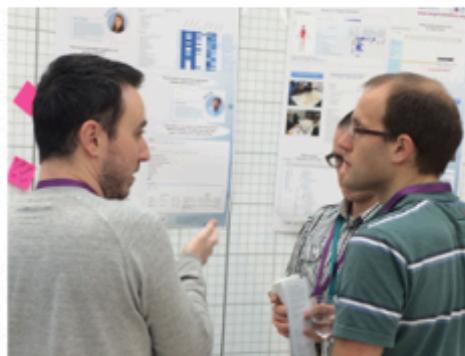
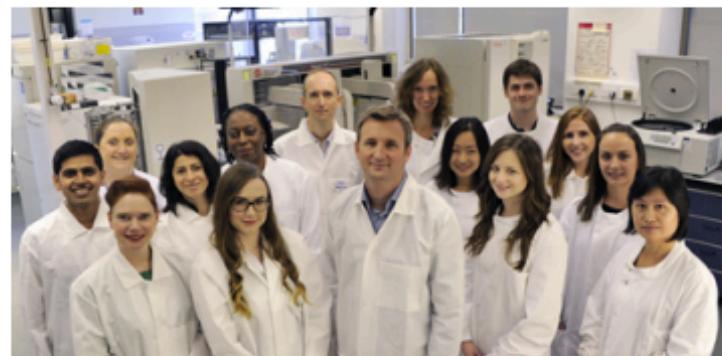
Open Targets Blog

blog.opentargets.org/



Open Targets

Acknowledgements



support@targetvalidation.org



Open Targets

How to cite us



The image shows the header of the Nucleic Acids Research journal website. At the top left is the "OXFORD JOURNALS" logo. The main title "Nucleic Acids Research" is displayed prominently in white text on a maroon background. Below the title are three navigation links: "ABOUT THIS JOURNAL", "CONTACT THIS JOURNAL", and "SUBSCRIPTIONS".

OXFORD JOURNALS

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS

Database issue

Koscielny, G. *et al.* (2017)

nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

Extras

Alternative ways to access the data

Looking for our entire datasets?

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

- All target-disease associations: 215 MB
- All evidence: 4.35 GB

Looking for extracts of our datasets?

- API: REST calls, Python client
- R client: maintained by the community



Open Targets

REST API endpoints



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

- Query association and evidence by gene identifiers and diseases
- Filter by type of evidence

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



Open Targets

GET

/public/association



Implementation notes

After integrating all evidence connecting a target to a specific disease, we compute an association score by mean of an harmonic sum. This association score provides an indication of how strong the evidence behind each connection is and can be used to rank genes in order of likelihood as drug targets. The association id is constructed by using the ensembl id of the gene and the EFO id for the disease (eg. ENSG00000073756-EFO_0003767). The method returns an association object, which contain data and summary on each evidence type included in the calculation of the score, as well as the score itself.

Parameters

Parameter	Value	Description	Parameter type	Data type
id	ENSG00000073756-EFO_0003767	an association ID usually in the form of TARGET_ID-DISEASE_ID	query	string

Response messages

HTTP status code	Reason	Model
200	Successful response	

[Try it out!](#)[Hide response](#)

Request URL

https://www.targetvalidation.org/api/latest/public/association?id=ENSG00000073756-EFO_0003767



Response body

```
{  
  "from": 0,  
  "facets": null,  
  "took": 6,  
  "therapeutic_areas": [],  
  "total": 1,  
  "data": [  
    {  
      "target": {  
        "gene_info": {  
          "symbol": "PTGS2",  
          "ensembl_id": "ENSG00000073756",  
          "name": "PTGS2",  
          "chromosome": 12, "start": 123456789, "end": 123456789},  
        "evidence": [{"source": "Ensembl", "score": 100, "type": "Gene-Disease"}, {"source": "OMIM", "score": 80, "type": "Gene-Disease"}],  
        "disease": {"id": "EFO_0003767", "name": "Prostaglandin synthase 2 deficiency", "ontology": "Disease", "category": "Phenotypic trait"},  
        "score": 100, "p_value": 0.001, "method": "Harmonic sum"}]  
  ]}
```

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

Python and R clients for the REST API

opentargets
latest

Search docs

Tutorial
High Level API
Low Level API
Code Documentation
Changelog

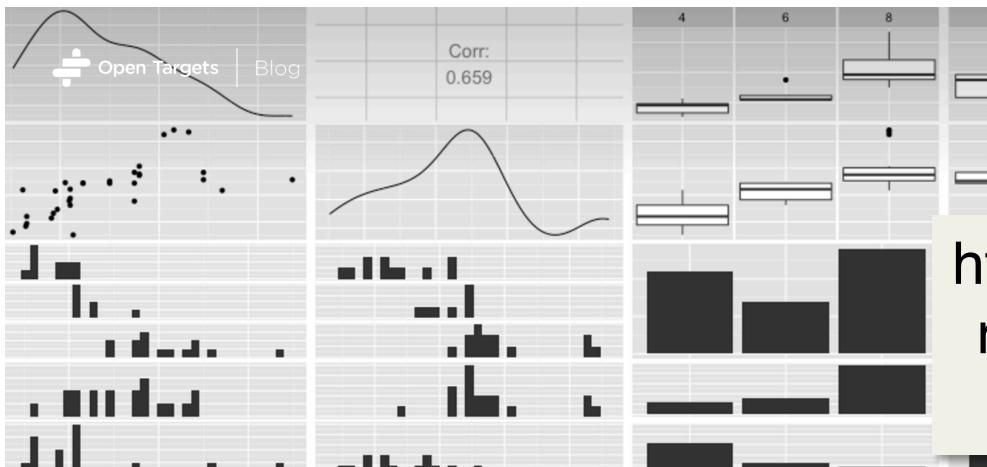
Docs » opentargets - Python client for targetvalidation.org

Edit on GitHub

opentargets - Python client for targetvalidation.org

opentargets is the official python client for the [Open Targets REST API](#) at [targetvalidation.org](#)

<http://opentargets.readthedocs.io>



[https://blog.opentargets.org/
rest-api-exploration-using-
an-r-client/](https://blog.opentargets.org/rest-api-exploration-using-an-r-client/)

How to access Open Targets
with R

Extras

From scoring an evidence...

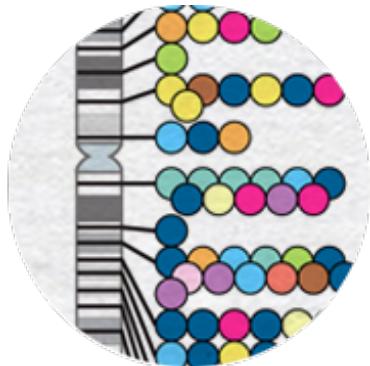
- $S = f * s * c$

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

- Evidence type: GWAS associations by lead SNP:



frequency = sample size (case versus control)

severity = predicted functional consequence

confidence = p -value reported in the paper

...through scoring a data source...

- 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

...to scoring a data type...

- For the association between *BRCA2* and breast carcinoma

Data type	Data sources
Somatic mutations	Cancer Gene Census, EVA, IntOgen

- Sort data sources scores
- Calculate the harmonic sum

...and getting the overall score

- Similar data sources are grouped for the score calculation
- Harmonic sum

Data type	Data sources
Genetic associations	GWAS catalog, UniProt, EVA, G2P
Somatic mutations	Cancer Gene Census, EVA, IntOgen
RNA expression	Expression Atlas
Drugs	ChEMBL
Affected pathways	Reactome
Text mining	Europe PMC
Animal models	PhenoDigm