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

exercises

answers

Open Targets

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

Open Targets

Course objectives

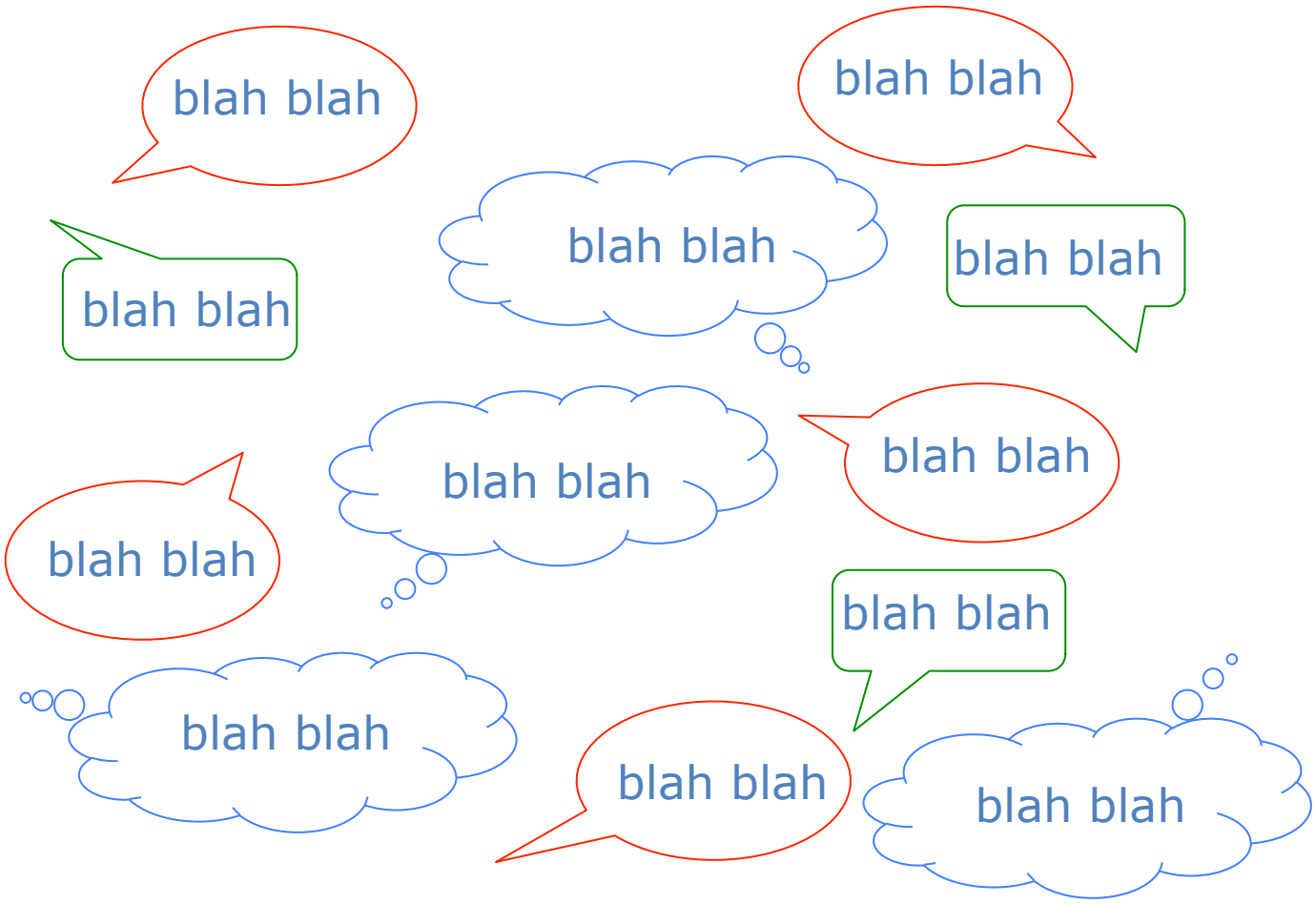
What is the **Open Targets Consortium**?

What is the **Open Targets Platform**?

How to **navigate** the Platform?

How to **connect** with us

Open Targets

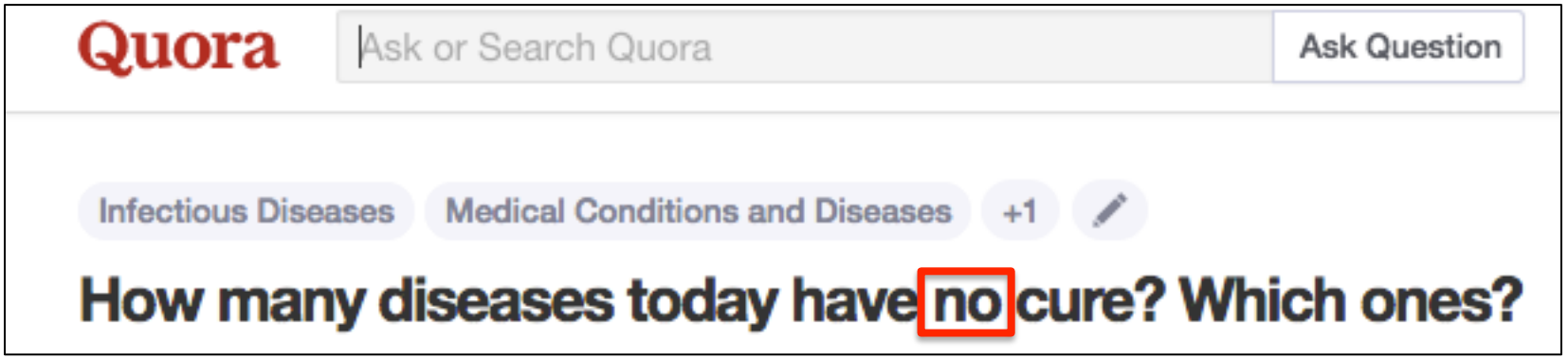


Open Targets

Do we need programmes for drug development and discovery?

❑ YES

❑ NO



Open Targets

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

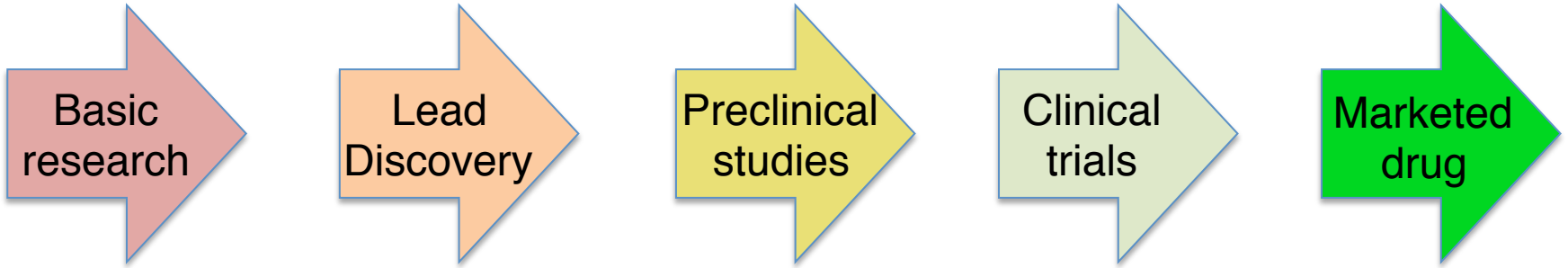
Ebola, HIV, Cancer, Dystonia are one of those disease which don't have any cure.

https://www.quora.com/How-many-diseases-today-have-no-cure-Which-ones-1

Open Targets

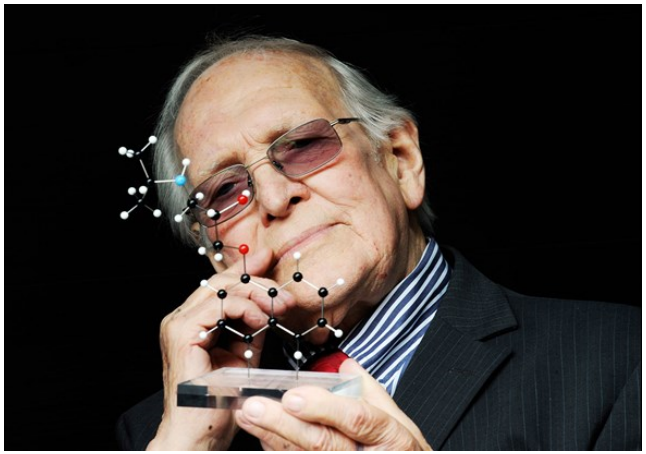
Drug development and discovery

- 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)

Open Targets



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

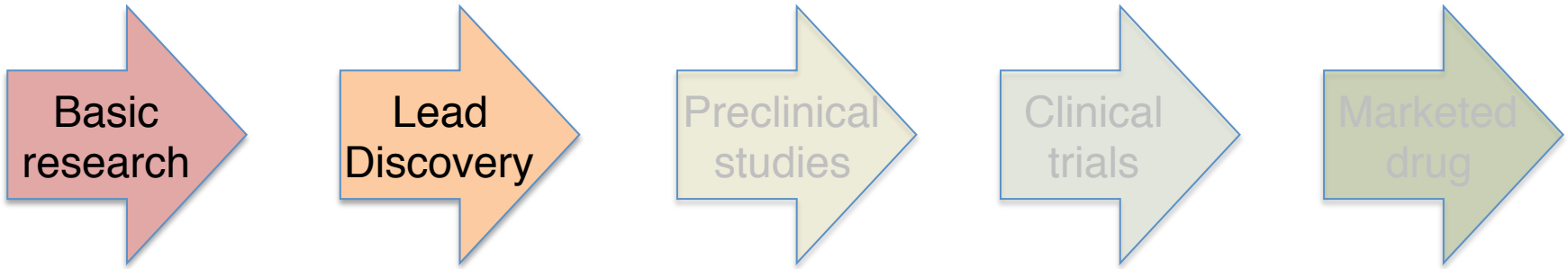
Open Targets

Drugs do fail in the clinic

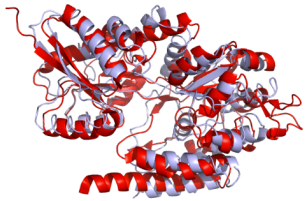
They don't work

They are not safe

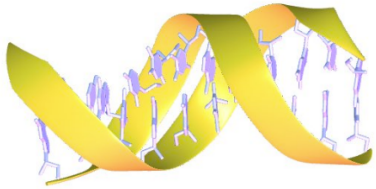
Important role of target identification and validation



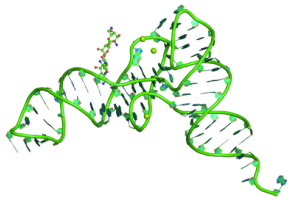
But what is a target?



protein



DNA

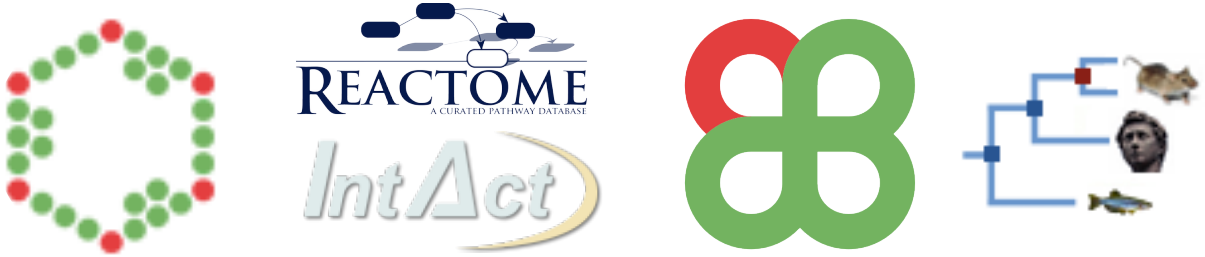


RNA

Open Targets

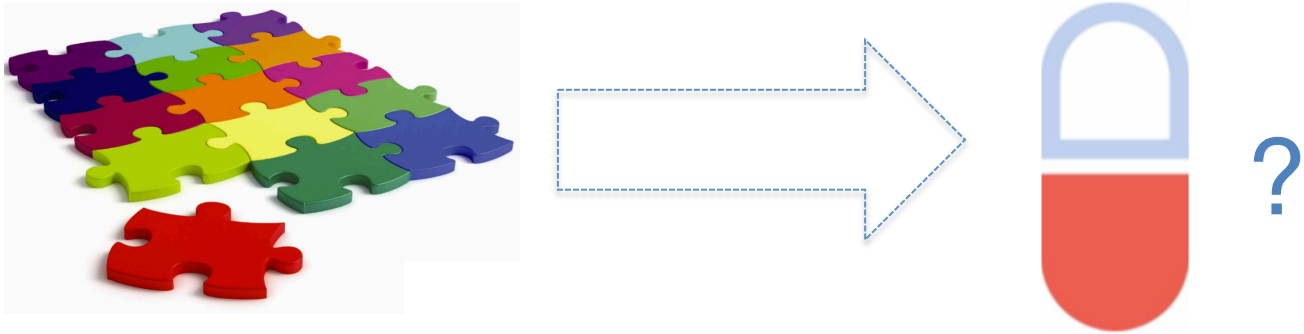
Public databases for drug discovery

- European Bioinformatics Institute (EBI-EMBL)



- Supporting all stages of drug discovery

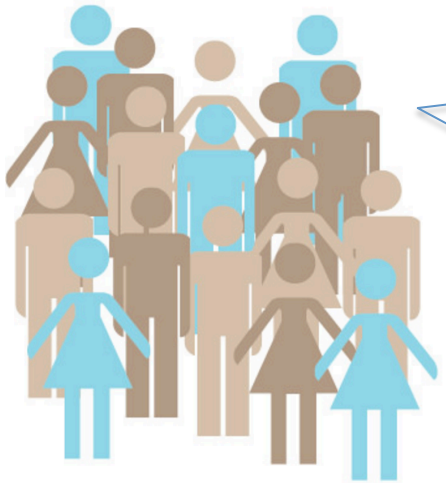
- ‘Fitting together like a jigsaw puzzle’ to build hypotheses



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

Open Targets

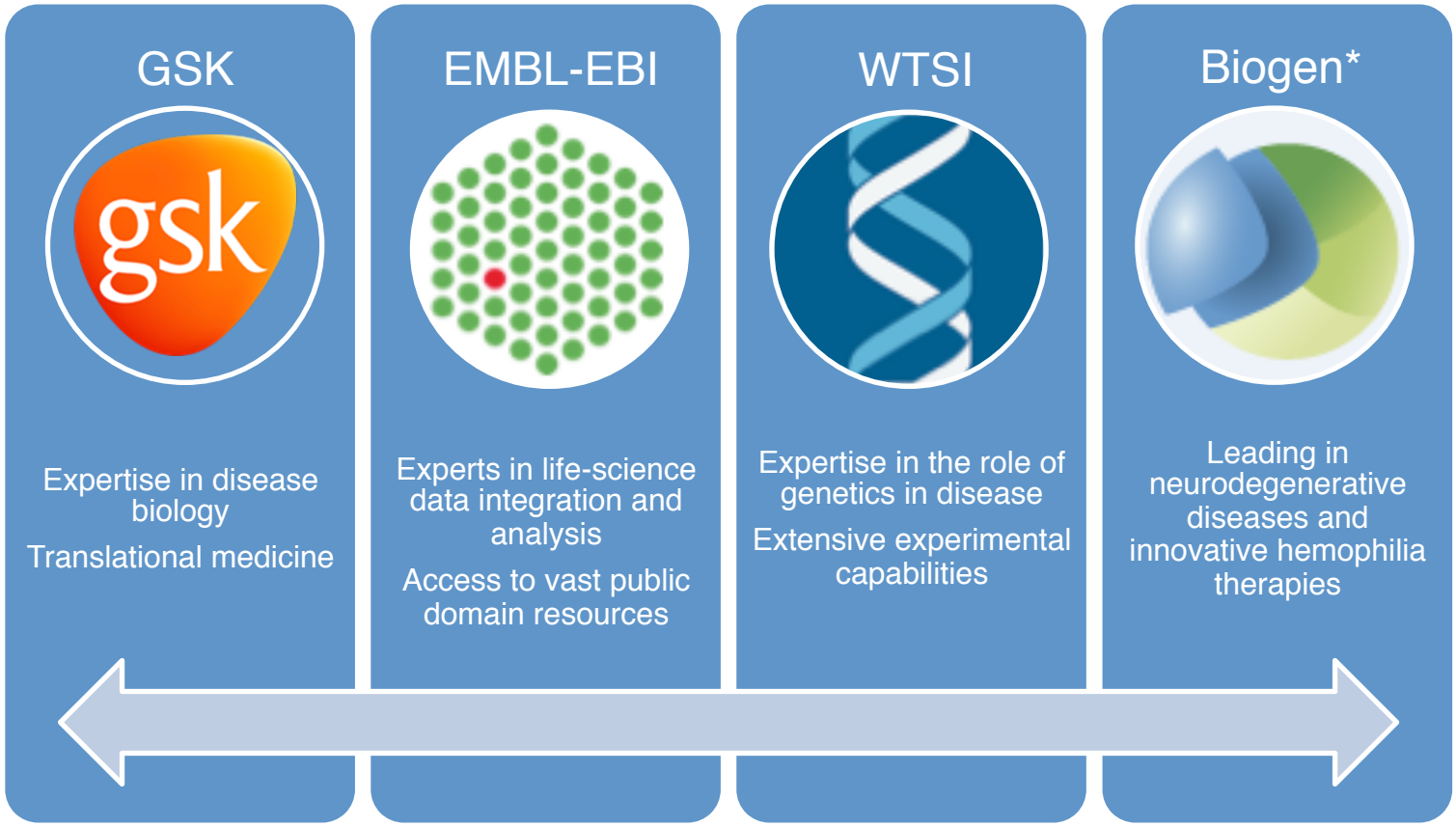
The consortium

- 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

Open Targets

Who is Open Targets?



*Biogen joined the consortium in February 2016

Open Targets

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

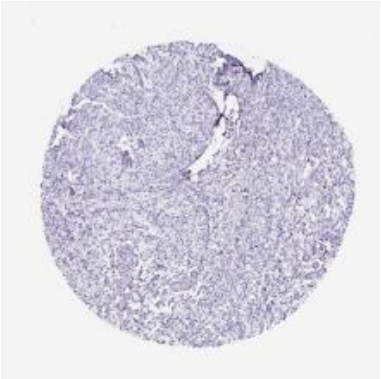
Generate new data
CRISPR
Organoids
Single cell RNASeq
Cell line fusion analyses
Metabolite GWAS

www.opentargets.org/projects

* Parts I and II are concurrent

Open Targets

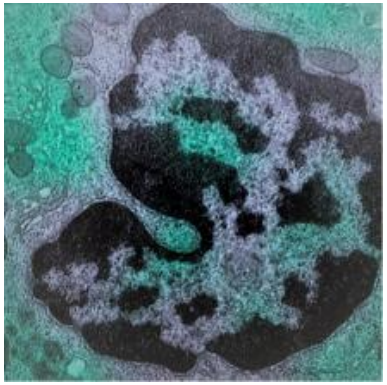
Experimental projects



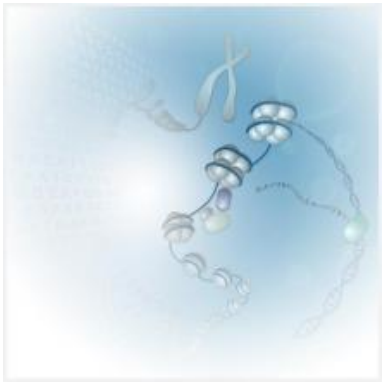
Oncology

Respiratory

IBD



Inflammation and immunity



Cell line epigenomes

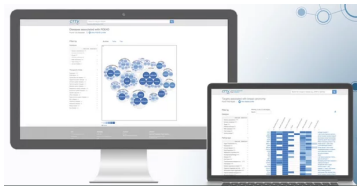
Neurodegenerative diseases

www.opentargets.org/projects

Open Targets

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

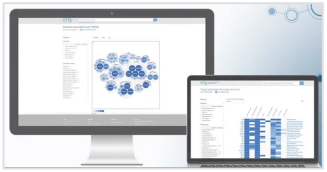
Can you explore Open Targets consortium website to find out:

- 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)

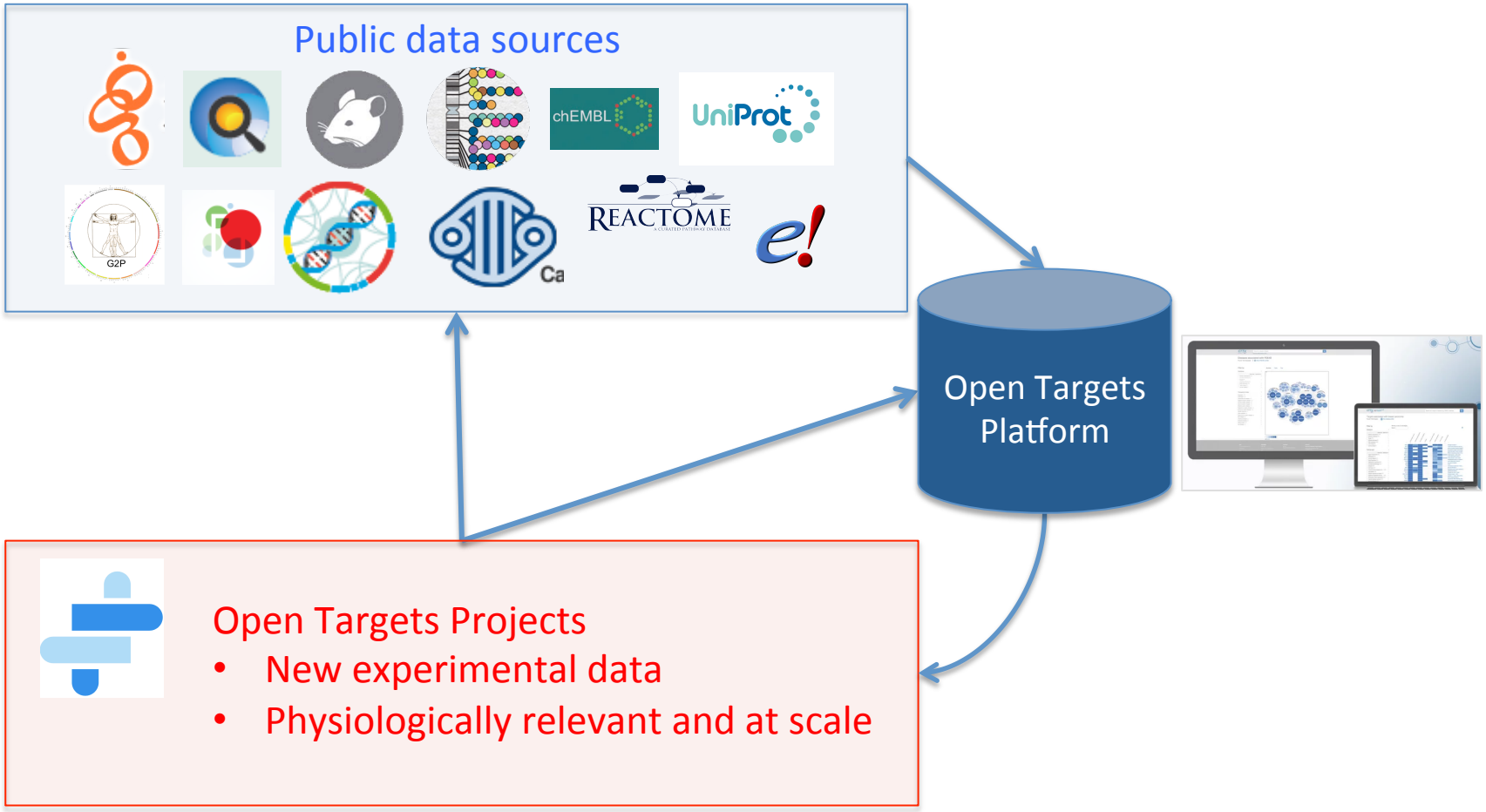
https://www.targetvalidation.org/



- Help scientists identify and prioritise relationships between targets and diseases

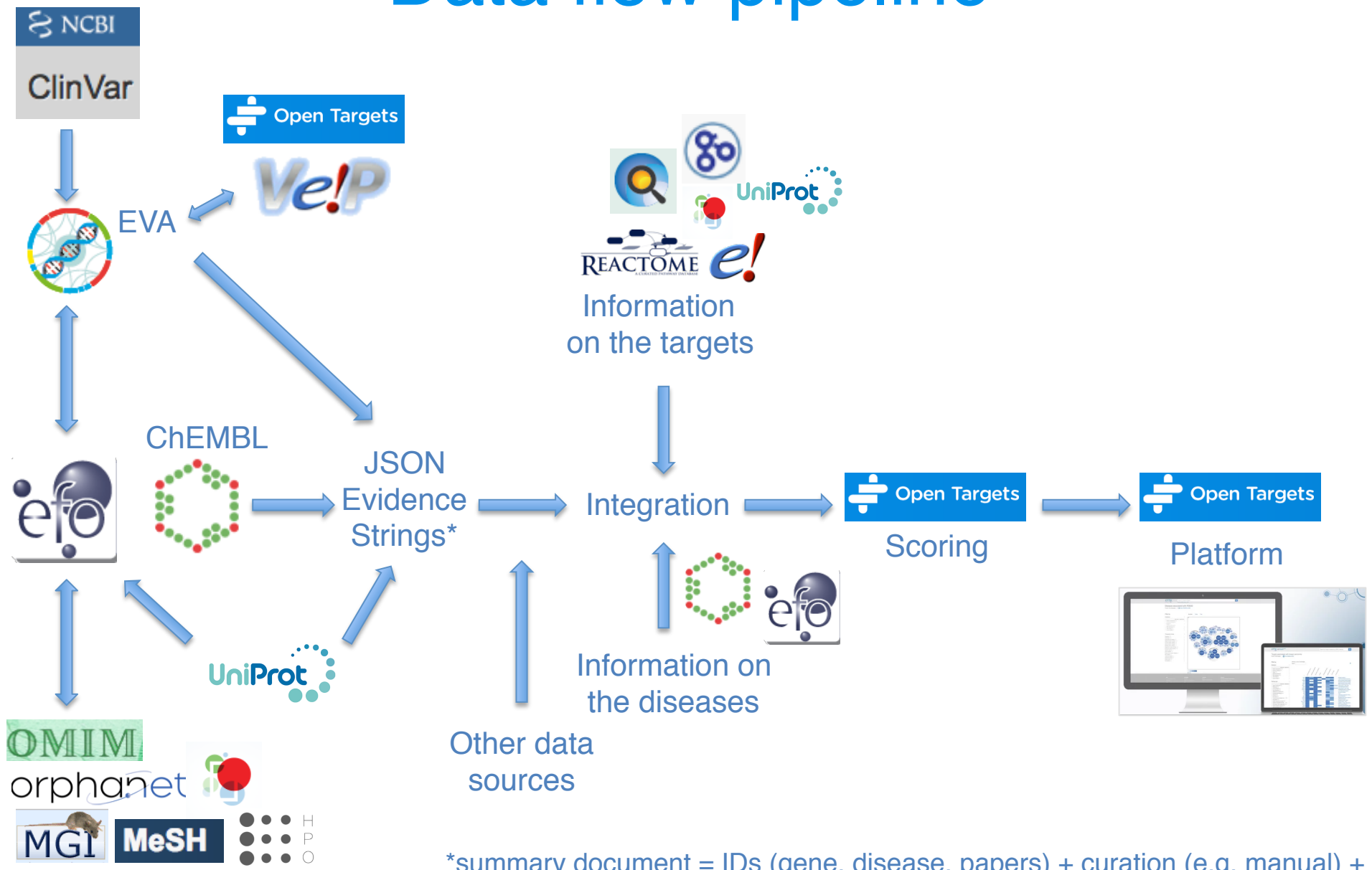
Open Targets

Platform to integrate existing and new data



https://www.targetvalidation.org/

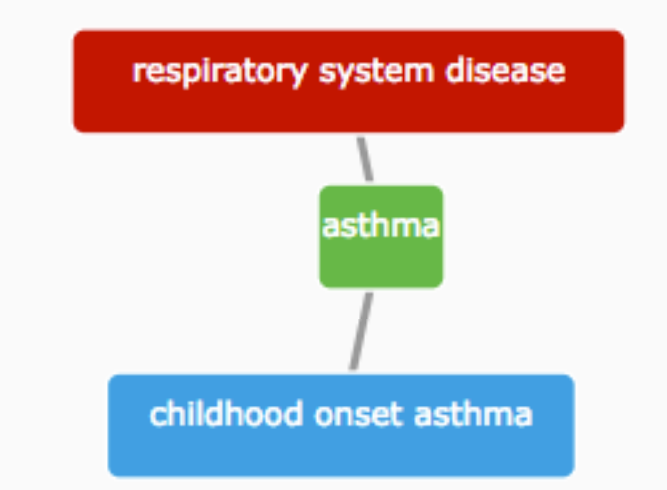
Data flow pipeline



*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

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

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

nature genetics

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journal ▾

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-Khaliq, 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

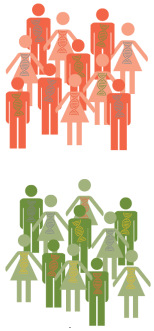
- Similar data sources are grouped for the calculation of the score

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

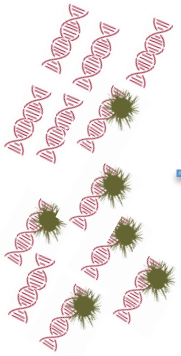
- Do you have a favourite data you would like us to have?

Data sources: GWAS catalog

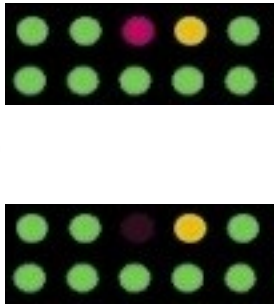
- GenomeWide Association Studies (www.nature.com/nrg/series/gwas/index.html)
- Array-based chips → genotyping 100,000 SNPs genomewide



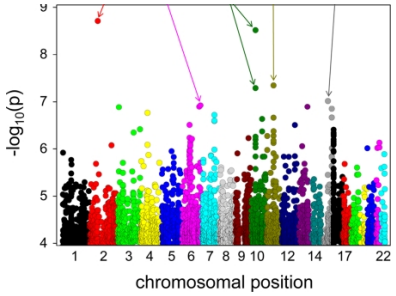
cases x controls



higher frequency of SNP in cases than controls



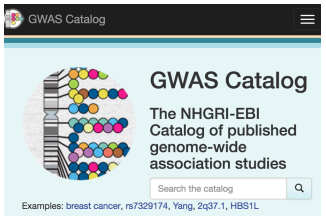
microarray chips



Manhattan plots

results published in papers

papers curated

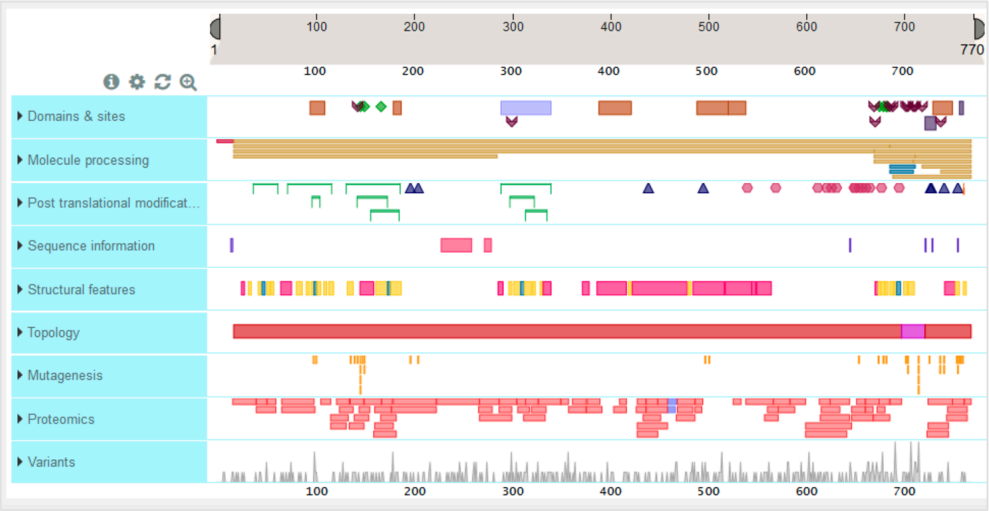


SNP-trait associations Published in the GWAS catalog

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

Data sources: UniProt* and UniProt literature**

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



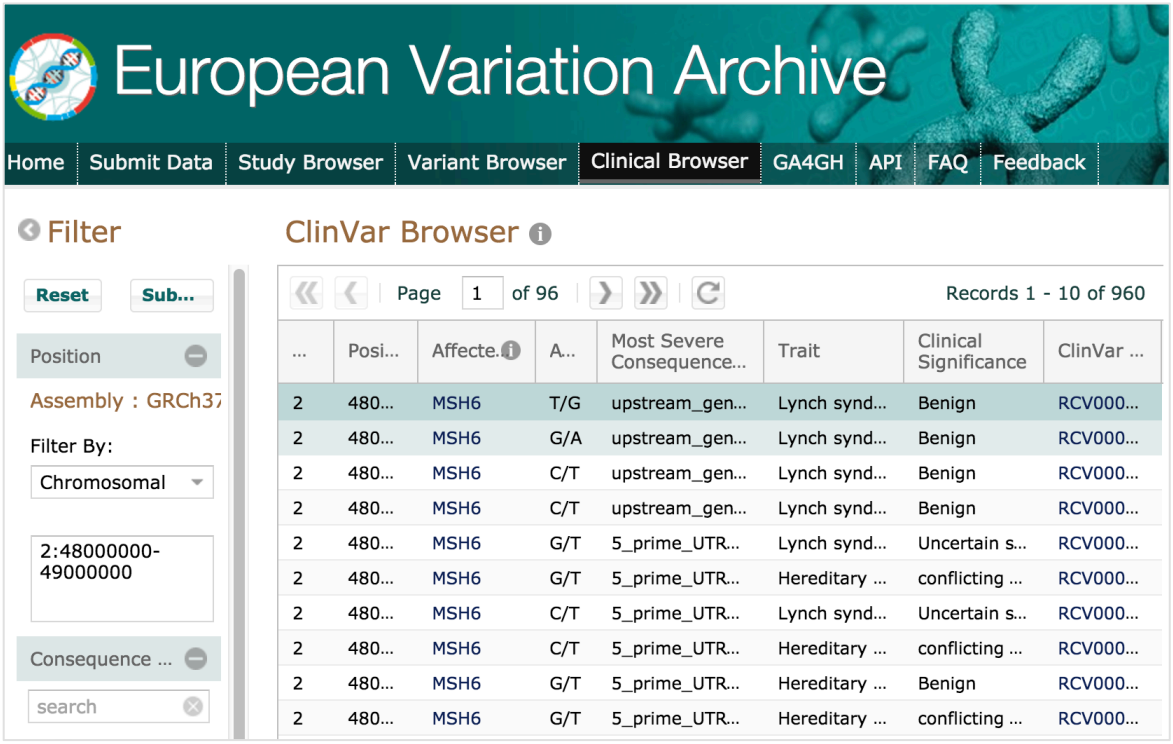
* Manual curation of variants in the coding region, seen in patients

** Associations between target and disease, no specific mutation

https://www.ebi.ac.uk/training/online/course/uniprot-quick-tourversion-0

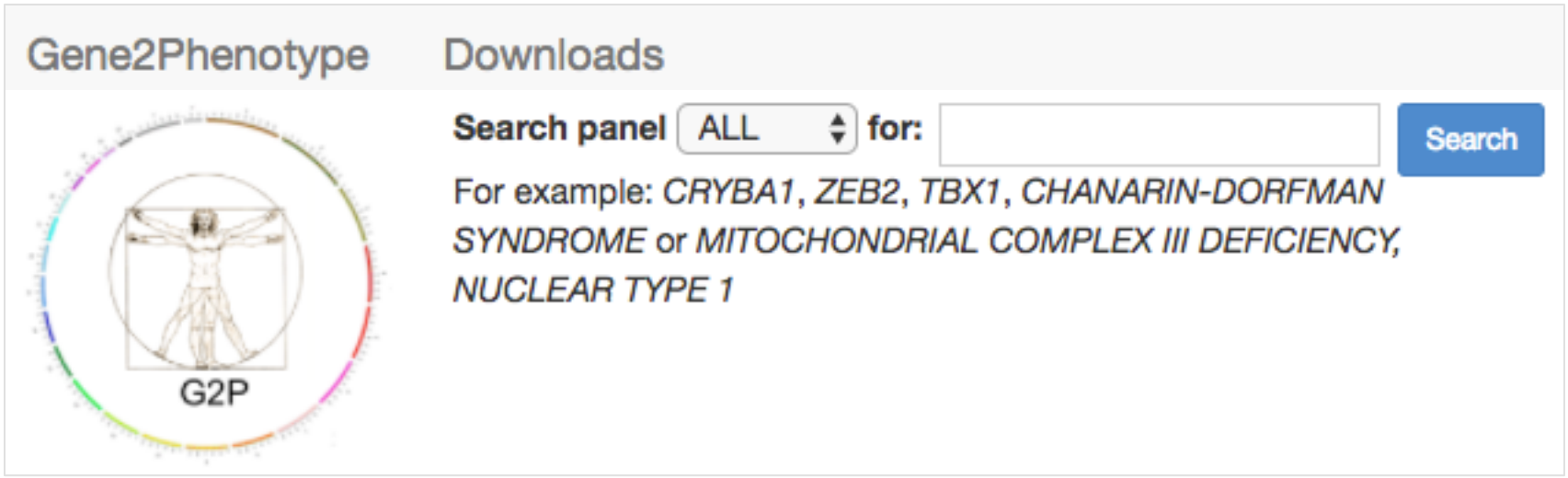
Data sources: EVA

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



http://www.ebi.ac.uk/training/online/course/european-variation-archive-embl-ebi-webinar?platform=hootsuite

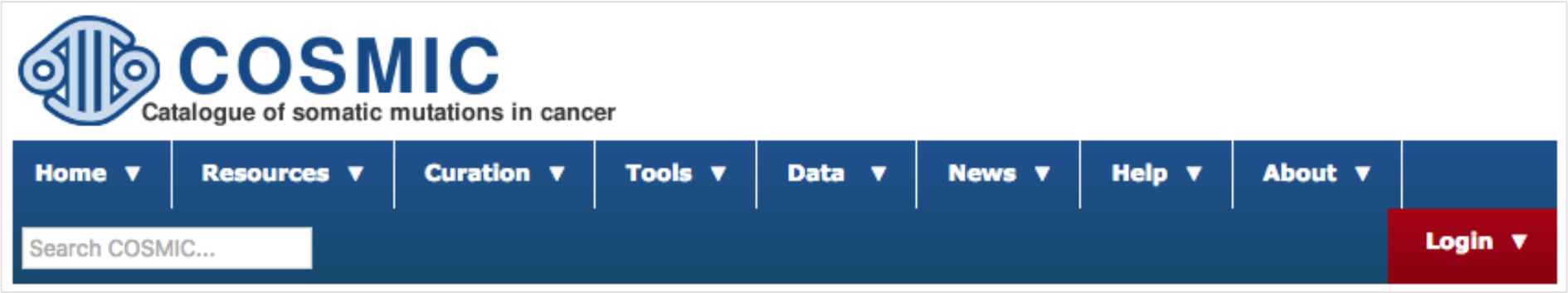
Data sources: Gene2Phenotype



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

https://www.ebi.ac.uk/gene2phenotype

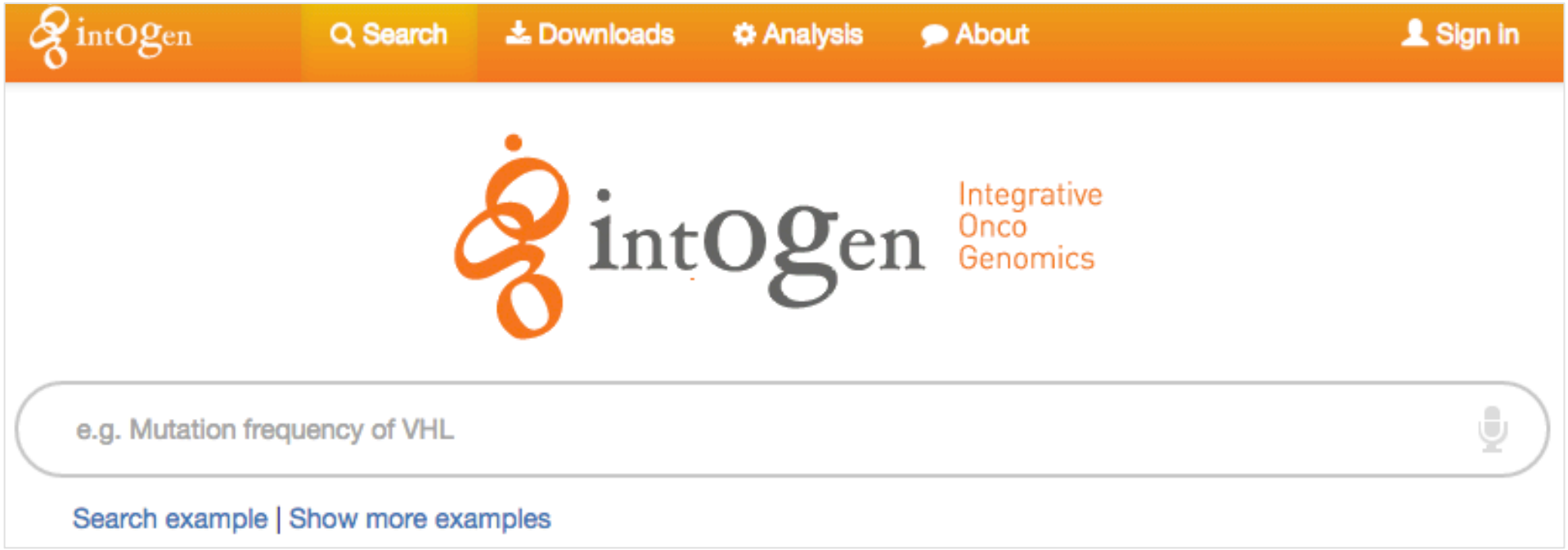
Data sources: The Cancer Gene Census



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

http://cancer.sanger.ac.uk/cosmic/census

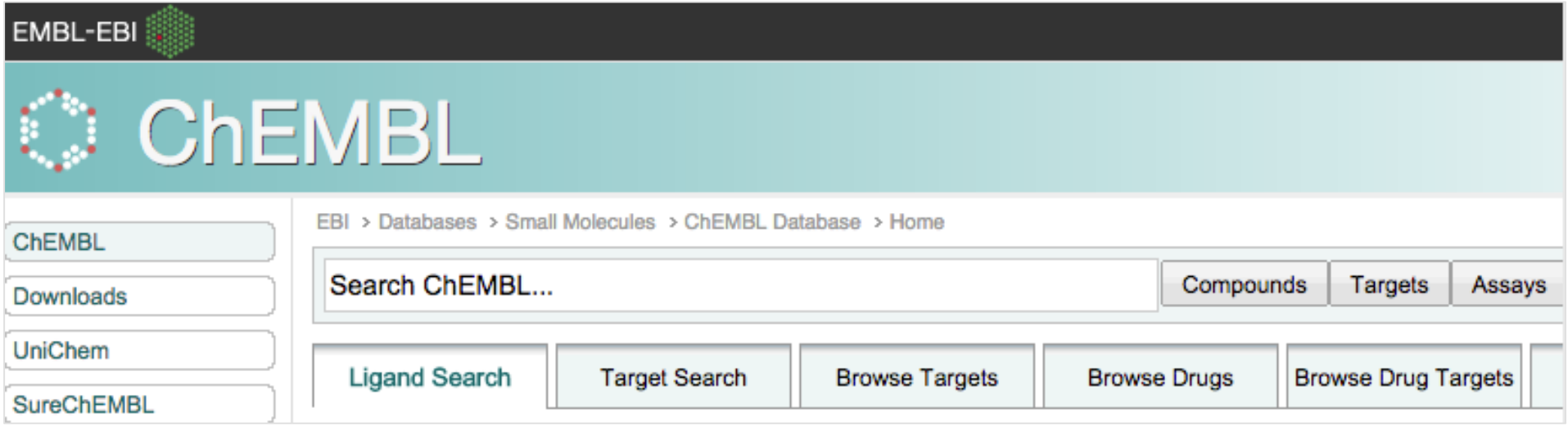
Data sources: IntOgen



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

https://www.intogen.org/search

Data sources: ChEMBL



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

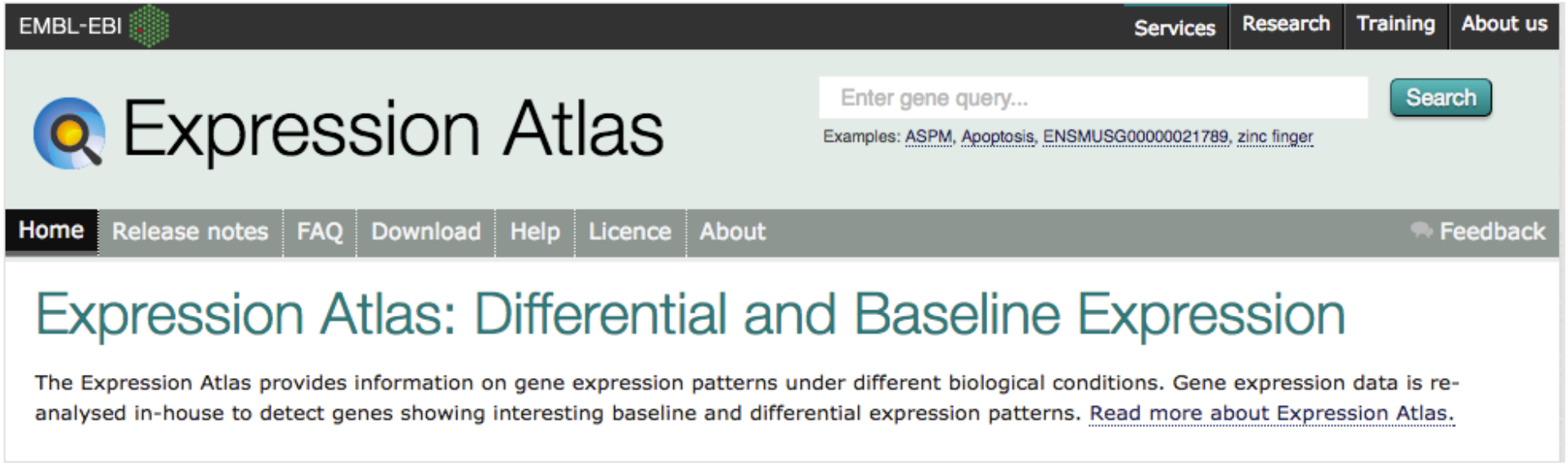
http://www.ebi.ac.uk/training/online/course/chembl-exploring-bioactive-drug-molecules

Data sources: Reactome

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

http://www.ebi.ac.uk/training/online/course/reactome-quick-tour

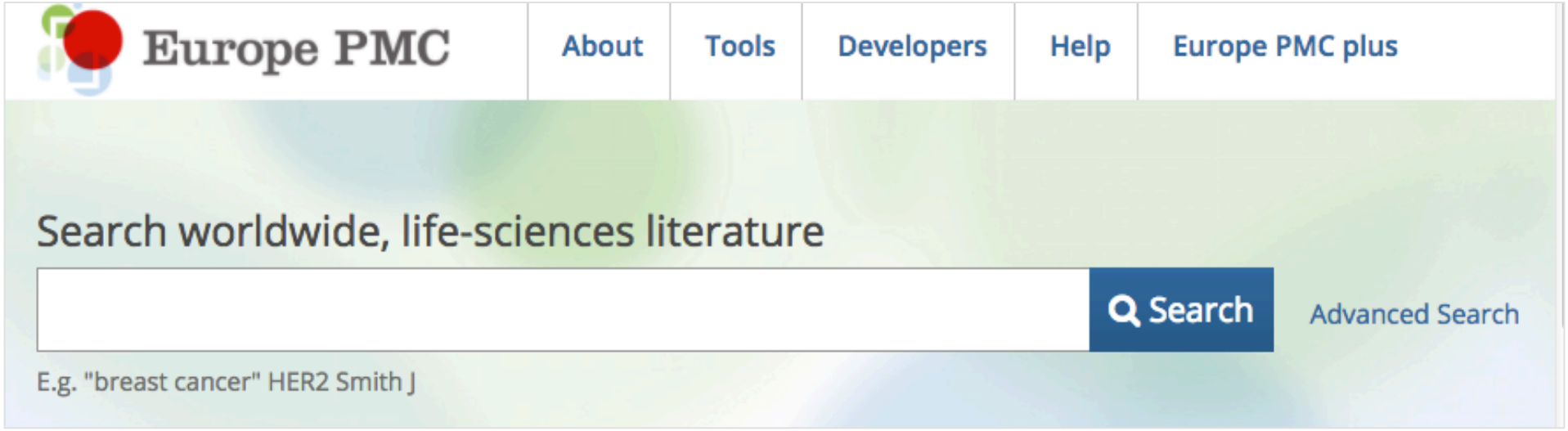
Data sources: Expression Atlas



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

http://www.ebi.ac.uk/training/online/course/expression-atlas-quick-tour-0

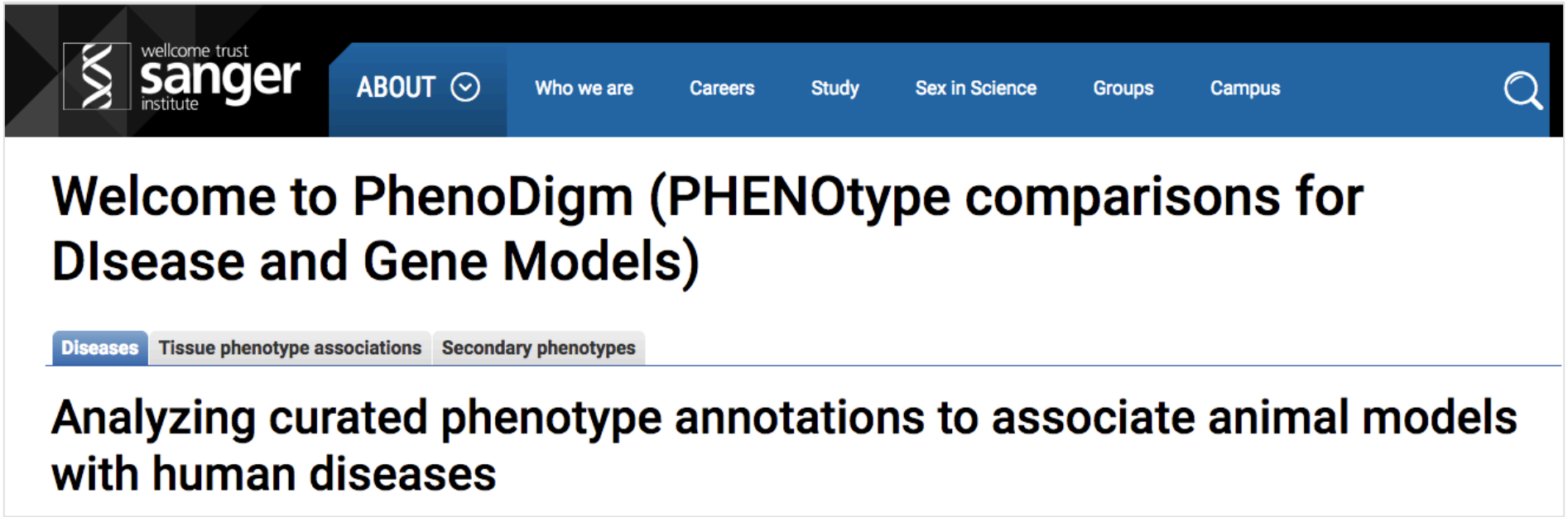
Data sources: Europe PMC



- 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)

http://www.ebi.ac.uk/training/online/course/europe-pmc-quick-tour-0

Data sources: PhenoDigm



- 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)

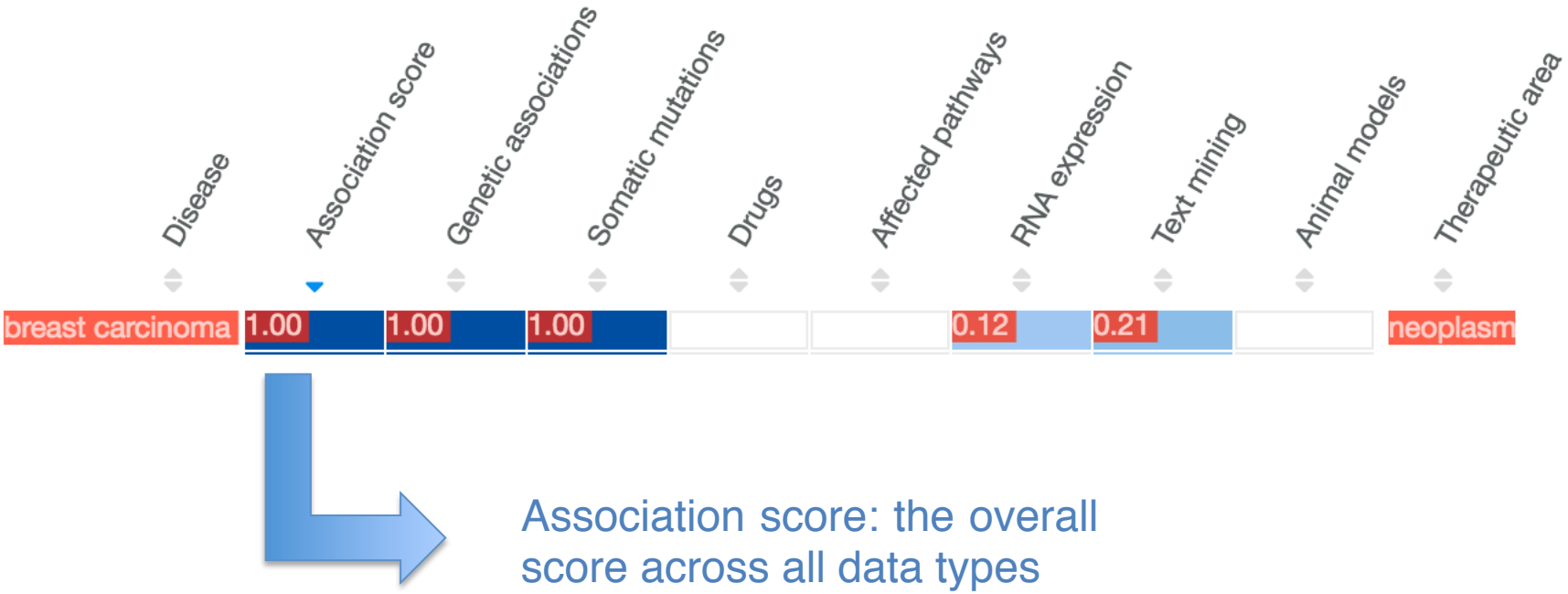
B) per data source (e.g. GWAS catalog)

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!

- 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

Open Targets

Wrap up

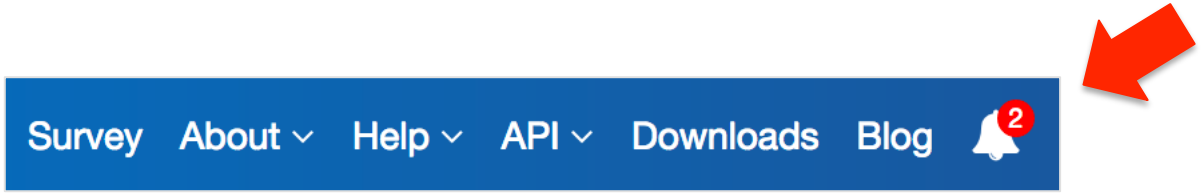
Open Targets Platform is the place:

- Identify and prioritise targets for drug discovery
- Target-disease associations: different sources
- Integrated information on target and diseases

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

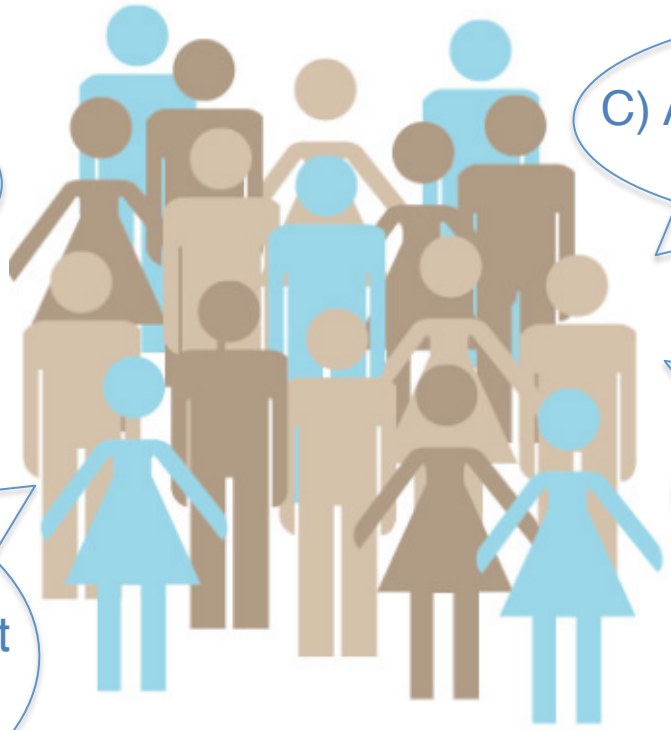
* https://www.targetvalidation.org/release-notes

Open Targets

We support decision-making

A) Which targets are associated with a disease?

B) What evidence supports this target-disease association?



C) Are there FDA drugs for this association?

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

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

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

Open Targets

Summary statistics

- Latest release: December 2016 *

31,071	2,559,080	8,659	13
targets	associations	diseases	data sources

Ensembl

efo Experimental Factor Ontology

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

Open Targets

> 2 million associations

Supported by publicly available data: the evidence

Text mining of **Europe PMC** abstracts and full text articles

RNA expression experiments from **Expression Atlas**

Mouse knockouts in disease models from **PhenoDigm**

Common and rare disease variants from **DECIPHER, GWAS catalog**, **UniProt, EVA**

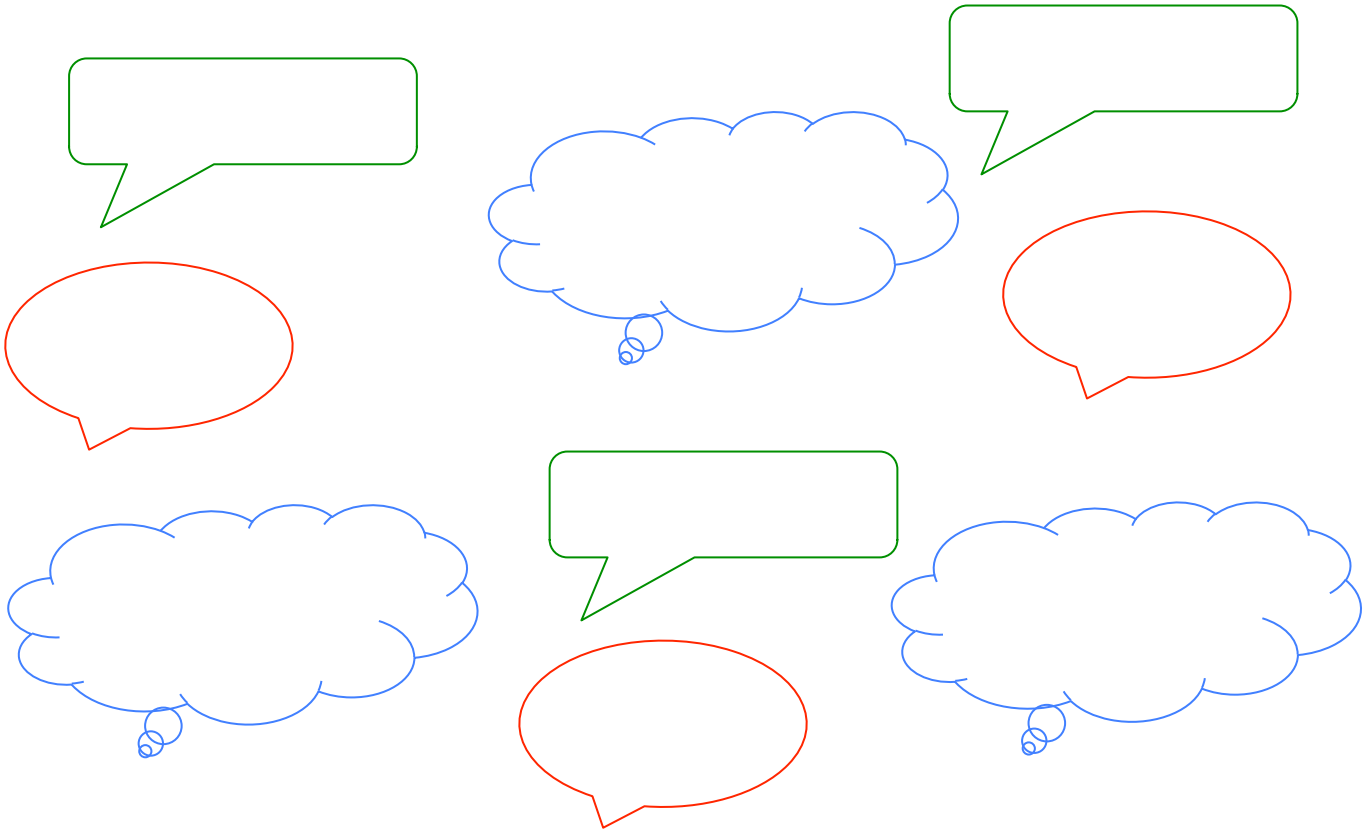
Pathways involved in disease associations from **Reactome**

Somatic mutations from **COSMIC** Cancer Gene Census and **IntOgen**

FDA-approved drugs from **ChEMBL**

Open Targets

Your take home message



Open Targets

Short feedback survey

http://tinyurl.com/cam-161216

Open Targets

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

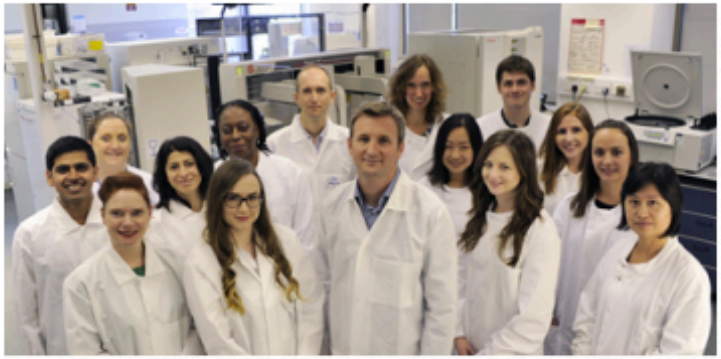
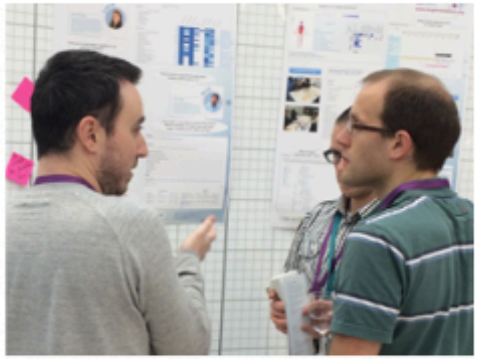
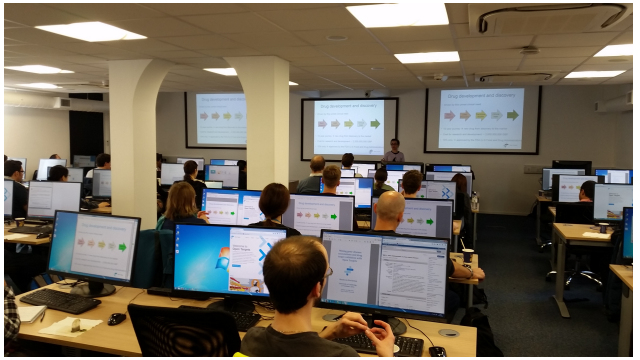
Open Targets Blog

blog.opentargets.org/

Open Targets

Acknowledgements



support@targetvalidation.org

Open Targets

How to cite us

Database issue

Koscielny, G. *et al.* (2017)
nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

Open Targets

Extras

Open Targets

Alternative ways to access the data

Looking for our entire datasets?

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

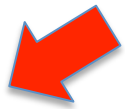
- All target-disease associations: 215 MB
- All evidence: 4.35 GB

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",
```


- 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)

jets

Python and R clients for the REST API

opentargets

latest

Search docs

Tutorial

High Level API

Low Level API

Code Documentation

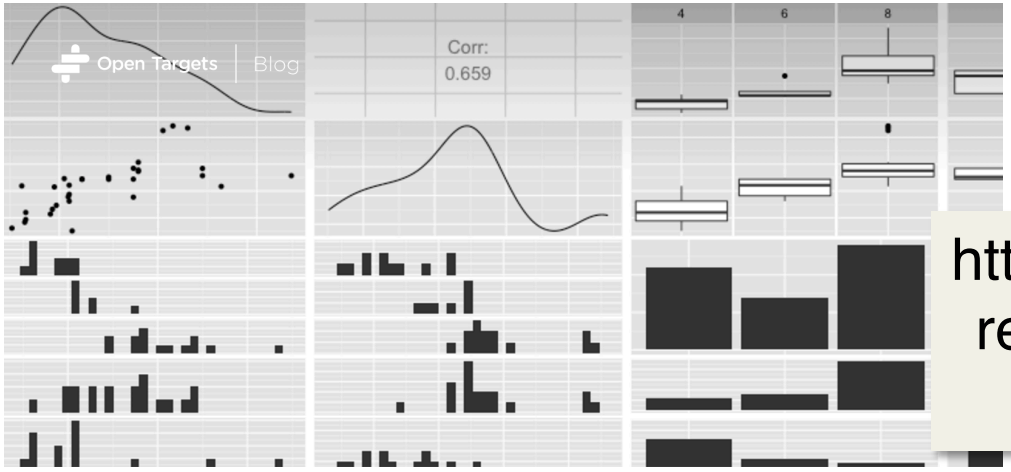
Docs » opentargets - Python client for targetvalidation.org

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/

How to access Open Targets with R

Extras

Open Targets

From scoring an evidence…

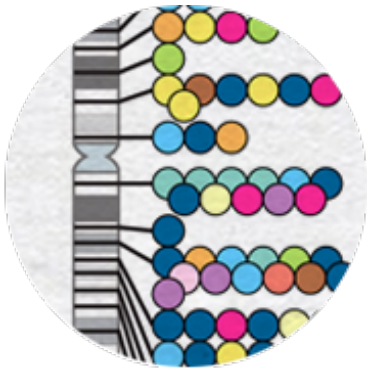
- S = f * s * c

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*

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

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

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

* PMID: 19107201, PMID: 20118918

…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