

Open Targets: integrating genetics, omics and chemistry for drug discovery



18th October 2018
Osong Medical Innovation Foundation
New Drug Development Center

Denise Carvalho-Silva

Open Targets / EMBL-EBI
United Kingdom

Notes

This booklet is based on the August 2018 release (18.08) of the Open Targets Platform. These are some useful links:

1) Science in Open Targets

<https://www.opentargets.org/science/>

2) Open Targets Platform documentation

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

3) Platform frequently asked questions

<https://docs.targetvalidation.org/faq/frequently-asked-questions>

4) Videos and animations

<https://tinyurl.com/opentargets-youtube>

5) Cite us

<https://academic.oup.com/nar/article/45/D1/D985/2605745>

6) Open Targets REST API docs

<https://api.opentargets.io/v3/platform/docs/swagger-ui>

7) GitHub

<https://github.com/opentargets>

Questions or suggestions?
support@targetvalidation.org

TABLE OF CONTENTS

OVERVIEW.....	4
INTRODUCTION TO OPEN TARGETS.....	5
HANDS-ON EXERCISES.....	31
Exercise 1: Vedolizumab and its targets	31
Exercise 2: Advancing research in the field of IBD	32
Exercise 3: Filtering Alzheimer's disease associations based on a list of targets	33
Exercise 4: LRRK2 in Parkinson's disease.....	34
EXTRA HANDS-ON EXERCISES	35
<i>Exercise E1: Assessing the specificity of targets for IBD</i>	<i>35</i>
<i>Exercise E2: Getting all disease associations and scores for three targets at once.....</i>	<i>35</i>
QUICK GUIDE TO DATABASES	37

OVERVIEW

Open Targets is a partnership to transform drug discovery through the systematic identification and prioritisation of targets.

We work to create a research and development (R&D) framework that can be applied to a wide range of human diseases. We will share our results openly with the scientific community.

The consortium was launched in March 2014 under the name of Centre for Therapeutic Open Targets (CTTV) and started with GlaxoSmithKline (<http://www.gsk.com/>), the Wellcome Sanger Institute (<http://www.sanger.ac.uk/>) and the EMBL-EBI (European Bioinformatics Institute) (<http://www.ebi.ac.uk/>). In February 2016, a Biogen (<https://www.biogen.com/>) joined the initiative. The consortium was rebranded to Open Targets in April 2016, and has welcome two new partners, Takeda in 2017 and Celgene in 2018.

In drug discovery, the *validation* of a target refers to the creation of a specific entity that modulates the activity of a target to provide therapeutic benefit to individuals with a disease.

The ultimate validation of a target is the creation of an effective therapeutic molecule. This is a long and costly endeavour with more high failure rates.

The goal of Open Targets is to transform this process by predicting if the modulation of a target is likely to provide therapeutic benefit. This would be done much earlier in the drug discovery process than is currently possible and far in advance of having a final, approved medicine.

Points covered in this workshop:

- The science carried out in Open Targets
- The Open Targets Platform
- How to browse the web interface of the Platform
- Alternative ways to access the Open Targets Platform data

INTRODUCTION TO OPEN TARGETS

Open Targets employs large-scale human genetics and genomics data to change the way drug targets are identified and prioritise. We have established a set of scientific projects to both integrate and generate data and analytical processes that implicate a target as valid.

Our experimental projects use CRISPR gene editing, induced pluripotent stem cells, single cell genomics, organoids to generate new data and provide insights in the validation of targets relevant to key therapeutic areas namely:

- Oncology
- Immunology
- Neurodegeneration

Our core bioinformatics and data pipelines team has developed the Open Targets Platform to provide easy access to data relevant to drug target identification and selection by a diverse audience of users.

More details on our projects can be found on our [Scientific Overview](#) page.

The Open Targets Platform

The Open Targets Platform is a web application that integrates and displays publicly available data to facilitate the identification and selection of targets for new therapies.

We use genetics, omics and chemical data from different [data sources](#) to associate genes and diseases. Similar data sources are combined into the following data types:

- Genetic associations
- Somatic mutations
- Drugs
- Affected pathways
- RNA expression

Text mining
Animal models.

The evidence (e.g. SNPs, scientific literature) is used to compute an [association score](#), which depends on the frequency of evidence, the confidence and severity (e.g. does the SNP change the amino acid of the target protein?). We then aggregate the evidence score using the sum of the [harmonic progression](#) to obtain the score at the data source and data type levels, as well as the overall score. The association score can be used to rank target and disease associations in the Platform.

The latest release of the Platform (August 2018) contains:

- 21,149 targets
- 10,101 diseases
- 2,920,121 associations between targets and diseases
- 6,507,752 evidence

The Open Targets Platform is for everyone from academia and industry. Our users can browse a target on a gene by gene (or disease by disease) basis, search for a list of targets in one go (up to 200 targets) using the batch search tool, carry out more complex queries using the REST-API, or download all evidence and association objects for downstream analyses.

What can you do with the Open Targets Platform?

- Find annotations for targets
- Find annotations for diseases
- Find which targets are associated with a disease
- Find which diseases are associated with a target
- Find the evidence supporting target-disease associations,

Connect with us

- ❖ [Open Targets Blog](#)
- ❖ Follow us on [Twitter](#), [Facebook](#), [LinkedIn](#), and [YouTube](#)

OPEN TARGETS PLATFORM: WALKTHROUGH

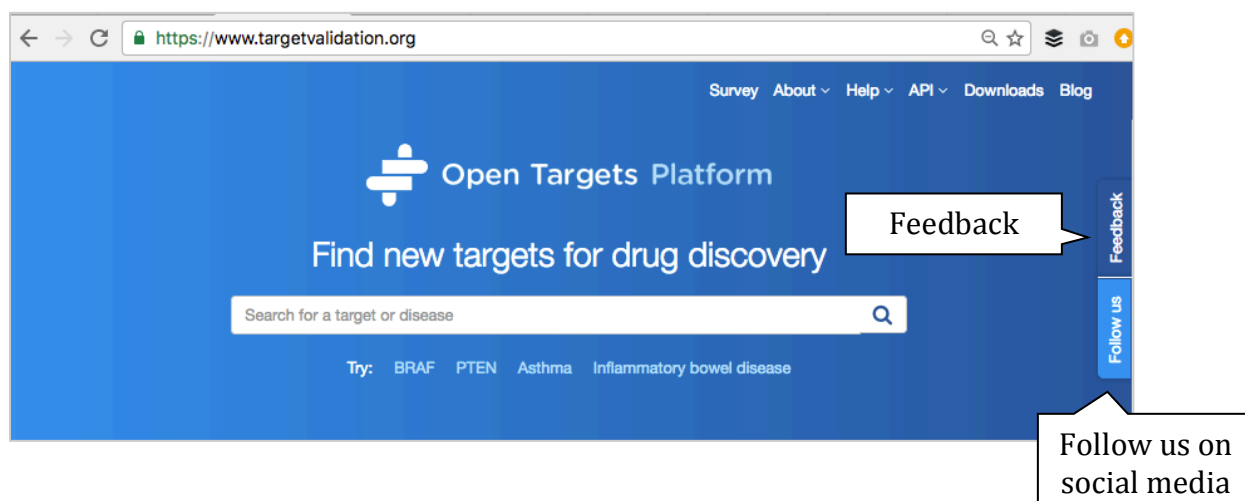
We will guide you through the website using multiple sclerosis (MS), as an example of a disease, then we will explore the evidence associating CD86 with MS.

The following points will be addressed during the walkthrough:

- How to find targets associated with multiple sclerosis
- How to filter down the number of targets based on specific type of evidence
- How to find out how strong the association between CD86 and MS is
- How to view the evidence that supports this association
- How to find other diseases (than MS) associated with CD86
- How to visualise CD86 in the context of the human genome
- How to find drugs currently in clinical trials for CD86
- How to filter the Open Targets associations by a list of genes already known to be linked to MS

Demo 1: Disease centric workflow

Go to www.targetvalidation.org and search for multiple sclerosis.



Select the first (best) hit:

multiple scler|

multiple sclerosis
2080 targets associated

Disease

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...

Targets
MBP myelin basic protein

Diseases
experimental autoimmune encephalomyelitis
[...] > central nervous system infe... > encephalomyelitis > experimental autoimmune e...
chronic progressive multiple sclerosis
autoimmune disease > multiple sclerosis > chronic progressive multipl...

You will see a page like this:

Total number of targets associated with multiple sclerosis

2678 targets associated with multiple sclerosis

Filter the results

Filter by

Data type

Genetic associations (372)
Somatic mutations (1)
Drugs (281)
Affected pathways (0)
RNA expression (995)
Text mining (1k)
Animal models (8)

Showing 1 to 50 of 2,678 targets

Search:

Download .csv

Target symbol	Overall association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
IL2RA									interleukin 2 receptor subunit ...
TNFRSF1A									TNF receptor superfamily me...
ITGA4									integrin subunit alpha 4
VDR									vitamin D receptor
KCNB2									potassium voltage-gated chan...
MS4A1									membrane spanning 4-domai...
S1PR1									sphingosine-1-phosphate rec...
CNR1									cannabinoid receptor 1
PTGS2									prostaglandin-endoperoxide s...
NR3C1									nuclear receptor subfamily 3 g...
IFNAR1									interferon alpha and beta rece...
ACHE									acetylcholinesterase (Cartwig...
CD52									CD52 molecule

Data types (Genetic Associations, Drugs, etc)

The current release of the Open Targets Platform (August 2018) lists 2678 targets associated with multiple sclerosis (MS).

The data types supporting these results are based on Genetic associations, Somatic mutations, Drugs, RNA expression, Text mining,

and Animal models. There is no evidence under Affected pathways available for this disease.

Check our help page to find out more about our data sources:
https://targetvalidation.org/data_sources.

The association table listing all > 2,000 targets associated with MS can be filtered depending on the following options:

A) Data types

- Genetic associations (e.g. GWAS catalog)
- Somatic mutations (e.g. Cancer Gene Census, EVA)
- Drugs (from ChEMBL)
- Affected Pathways (i.e Reactome, SLAPenrich, PROGENy)
- RNA expression (from Expression Atlas)
- Text mining (from EuropePMC)
- Animal models (from PhenoDigm)

B) Pathway types

- Immune System
- Signal Transduction
- Metabolism
- ...

C) Target class

- Enzyme
- Membrane receptor
- ...

D) RNA tissue specificity

Select the organs (or anatomical system) where the target is significantly more expressed in the selected tissues than the mean of the other tissues

E) Your target list

Upload your own list of genes (in official gene symbols from HGNC or Ensembl Gene IDs)

Data types: we collect data from various sources and combine them into categories called Data types. Example of data sources are GWAS catalog and UniProt, both combined into Data types. Note that data from an individual source can contribute to different Data types, e.g.

data from EVA is observed in two data types, Genetic associations and Somatic mutations.

Pathway types: Reactome is the resource that provides us with pathway classification e.g. immune system (and its subtypes e.g. neutrophil degranulation), hemostasis (platelet degranulation), etc.

<http://www.reactome.org/>

Target class: ChEMBL provides us targets grouped into different classes such as Enzyme, Ion channel, etc.

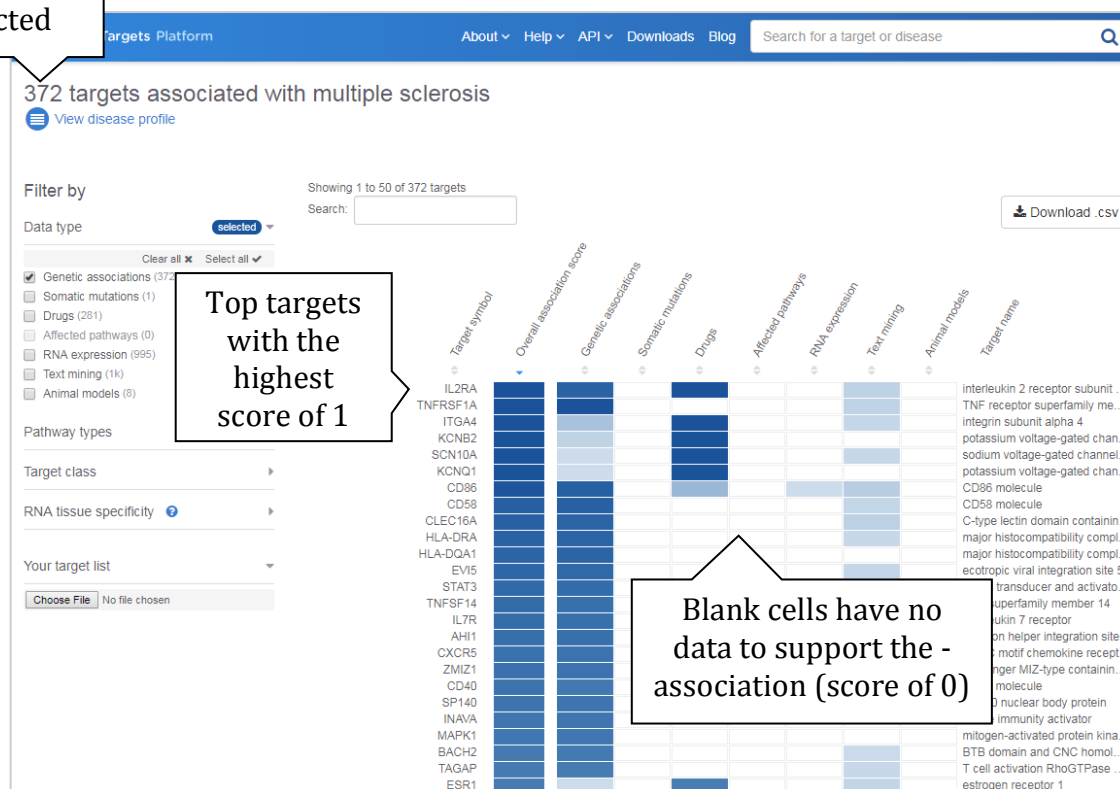
<https://www.ebi.ac.uk/chembl/>

RNA tissue specificity: RNA tissue specificity: the tissue specificity of a target is computed as the number of standard deviations from the mean of the log RNA expression of the target across the available tissues. This is a standard z-score calculation. A target is considered to be tissue specific if the z-score is greater than 0.674 (or the 75th percentile of a perfect normal distribution). We remove data for under-expressed targets before the z-score calculation. This RNA expression data comes from Expression Atlas.

Your target list: you can upload a list of targets for any given disease to restrict the table to show only the targets in your list, to help you to see the evidence Open Targets has integrated for them.

Let's now apply the 'Genetic associations' filter. The number of targets goes down to 372:

Genetic
associations
selected



These are targets associated with MS based on genetic variants (germline) only. Genetic variants are SNPs (single nucleotide polymorphisms but can also be mutations, or short indels) from data sources such as the GWAS Catalog, UniProt, PheWAS catalog, Genomics England PanelAPP. They are associated with either common or rare diseases.

The table is sorted by default with the best hit at the top of the table i.e. IL2RA. The best hit is the target that contains the highest (overall) association score. The association score can vary from 0 to 1. We sum up the scores from different data sources to obtain the overall association score. Because of this harmonic sum, the score can be higher than 1, but we will always cap it to 1.

Different weight is given to different data types when we compute the score. RNA expression, animal models and text mining data is downweighted by a factor of 0.5 (Expression Atlas) and 0.2 (both PhenoDigm and Text mining from Europe PMC abstracts and full text research papers).

You can sort the table by alphabetical order of the list of targets, or by the association score values (either overall or per data type e.g.

Genetic associations, Drugs, Text mining, etc). The association score varies from 0 to 1, the closer to 1 the more evidence we have for an association. This score is computed for each piece of evidence that is used to support the association. Individual scores within data sources and data types are combined to give the overall score - 'Overall association score' column in the table below:

Showing 1 to 50 of 372 targets

Search:

[Download .csv](#)

Target symbol	Overall association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
IL2RA									interleukin 2 receptor subunit ...
TNFRSF1A									TNF receptor superfamily me...
ITGA4									integrin subunit alpha 4
KCNB2									potassium voltage-gated chan...
SCN10A									sodium voltage-gated channel...
KCNQ1									potassium voltage-gated chan...
CD86									CD86 molecule
CD58									CD58 molecule
CLEC16A									C-type lectin domain containin...
HLA-DRA									major histocompatibility compl...
HLA-DQA1									major histocompatibility compl...

Click here to sort the column by alphabetical order of the gene symbols

Click on the arrows to sort the rows by (increased or decreased) score values of individual data types.

Let's now explore the evidence used to associate CD86 and multiple sclerosis.

Click on any of the columns in the CD86 row in the table below:

Showing 1 to 50 of 372 targets

Search:

[Download .csv](#)

Target symbol	Overall association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
IL2RA									interleukin 2 receptor subunit ...
TNFRSF1A									TNF receptor superfamily me...
ITGA4									integrin subunit alpha 4
KCNB2									potassium voltage-gated chan...
SCN10A									sodium voltage-gated channel...
KCNQ1									potassium voltage-gated chan...
CD86									CD86 molecule
CD58									CD58 molecule
CLEC16A									C-type lectin domain containin...
HLA-DRA									major histocompatibility compl...
HLA-DQA1									major histocompatibility compl...

Link to evidence page for CD86

You will land in the evidence page for the association between CD86 and multiple sclerosis:



The coloured petals on the flower plot represent the data types that support this association namely:

Genetic associations

Drugs

RNA expression

Text mining

Note: If you wish to suggest data types or resources we could/should incorporate into our Platform, please get in touch:
support@targetvalidation.org

Grey areas in the flower plot above indicate there is no information for those data types. There are no somatic mutations, affected pathway information, and animal models to support the association.

Let's now scroll down on the page and expand the tabs available. In this example, the genetic evidence comes solely from the GWAS Catalog. These are risk-associated variants, some of them available in dbSNP (the hint is on the rs IDs, i.e. rs9282641, rs4308217, rs2255214, and rs2681424), others may be provided in the genomic coordinates format only.

Genetic associations

Table **Browser**

Common diseases
Source: [GWAS catalog](#), [PheWAS catalog](#)

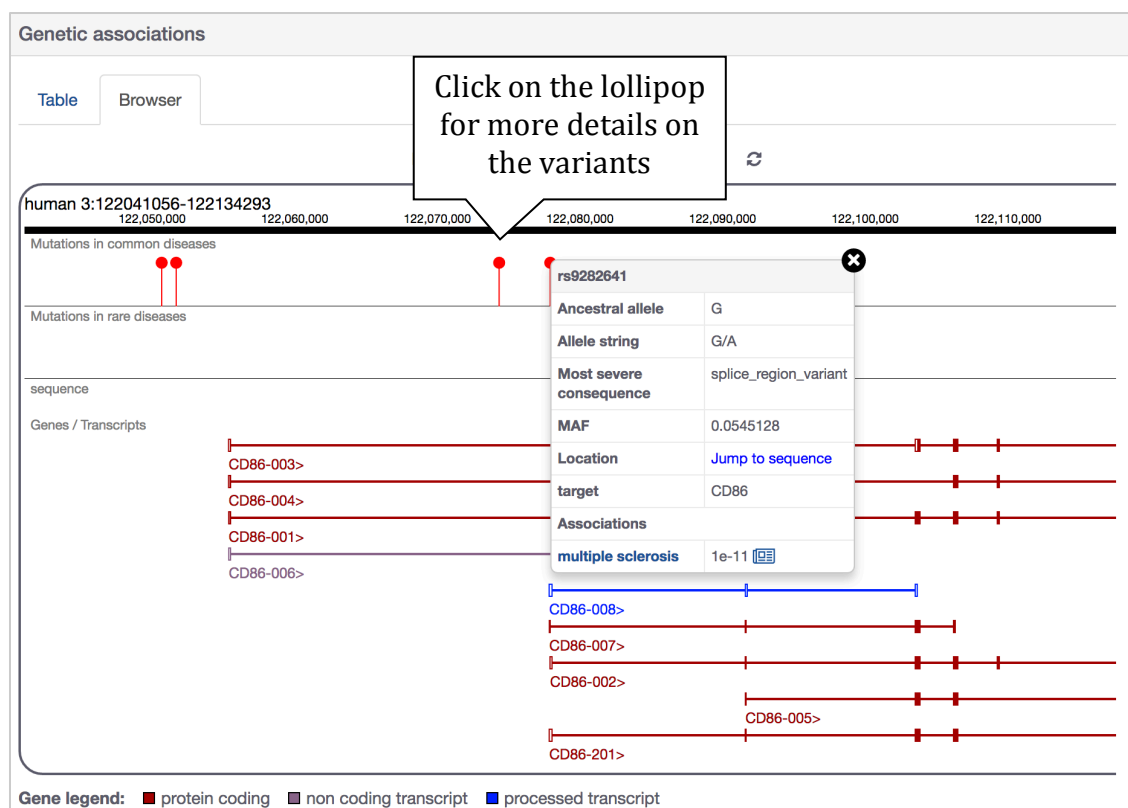
Showing 1 to 5 of 5 entries
Search:

Disease	Variant	Variant type	Evidence source	p-value	Publications
multiple sclerosis	rs9282641	splice region variant	gwas catalog	1e-11	1 publication
multiple sclerosis	rs2255214	upstream gene variant	gwas catalog	1e-24	1 publication
multiple sclerosis	rs4308217	intron variant	gwas catalog	6e-8	1 publication
multiple sclerosis	rs2255214	upstream gene variant	gwas catalog	5e-8	1 publication
multiple sclerosis	rs2681424	upstream gene variant	gwas catalog	2e-7	1 publication

Show 10 entries

Previous **1** Next

In addition to the table above, you can also explore the variants in a genome-context view. Click on the Browser tab:



This view is interactive: you can zoom in and out, scroll along the genome and find out more about the gene (s), transcript (s), and the genetic variants (depicted as lollipops).

Note: The assembly we use is GRCh38, also known as hg38.

Let's now expand the 'Drugs' tab to find out if there are drugs currently in clinical trials targeting CD86 in patients with multiple sclerosis. There is just one drug (i.e. ABATACEPT) on phase II and I of different clinical trials. There are three studies from clinicaltrials.gov, NCT01116427 and NCT00076934 (both for relapsing-remitting multiple sclerosis), and NCT00035529 (for multiple sclerosis).

Still on the same page, you can also to find out the research articles that have been mined for the co-occurrence of gene name and disease in the same sentence (under Text mining):

Text mining

Source: Europe PMC


Shown are the 15 articles where **target** and **disease** are found in the same sentence.

Showing 1 to 10 of 15 entries
Download .tsv

Disease	Publication	Year
multiple sclerosis	<p>Multiple sclerosis risk variants alter expression of co-stimulatory genes in B cells. Smets I <i>et al.</i> Brain undefined(undefined):undefined PMID: 29361022</p> <p>Abstract</p> <p>The increasing evidence supporting a role for B cells in the pathogenesis of multiple sclerosis prompted us to investigate the influence of known susceptibility variants on the surface expression of co-stimulatory molecules in these cells. Using flow cytometry we measured surface expression of CD40 and CD86 in B cells from 68 patients and 162 healthy controls that were genotyped for the multiple sclerosis associated single nucleotide polymorphisms (SNPs) rs4810485, which maps within the CD40 gene, and rs9282641, which maps within the CD86 gene. We found that carrying the risk allele rs4810485T lowered the cell-surface expression of CD40 in all tested B cell subtypes (in total B cells $P = 5.10 \times 10^{-5}$ in patients and 4.09×10^{-6} in controls), while carrying the risk allele rs9282641G increased the expression of CD86, with this effect primarily seen in the naive B cell subset ($P = 0.043$ in patients and 5.38×10^{-5} in controls). In concordance with these results, analysis of RNA expression demonstrated that the risk allele rs4810485T resulted in lower total CD40 expression ($P = 0.057$) but with an increased proportion of alternative splice-forms leading to decoy receptors ($P = 4.00 \times 10^{-7}$). Finally, we also observed that the risk allele rs4810485T was associated with decreased levels of interleukin-10 ($P = 0.020$), which is considered to have an immunoregulatory function downstream of CD40. Given the importance of these co-stimulatory molecules in determining the immune reaction that appears in response to antigen our data suggest that B cells might have an important antigen presentation and immunoregulatory role in the pathogenesis of multiple sclerosis.</p> <p>Introduction and background: 2 matched sentences</p> <p>Results: 1 matched sentence</p> <p>Discussion: 7 matched sentences</p> <p>Other: 4 matched sentences</p> <p>Figure: 3 matched sentences</p>	2018
multiple sclerosis	<p>Polymorphisms of RPS6KB1 and CD86 associates with susceptibility to multiple sclerosis in Iranian population. Abdollah Zadeh R <i>et al.</i> Neurol Res 39(3):217-222 PMID: 28079472</p> <p>Abstract</p> <p>Multiple sclerosis (MS) is the most prevalent disorder of nervous system inflammation which involves demyelination of spinal cord; this process depends on both environmental and genetic susceptibility factors. In the present study, we examined the association between two SNPs in RPS6KB1 (rs180515) and CD86 (rs9282641) with MS in Iranian population. RPS6KB1 gene encodes p70S6K1 protein which plays a key role in mTOR signaling pathway, while CD86 gene codes a membrane protein type I which belongs to immunoglobulin super family act on co-stimulation signaling pathway. In this case-control study 130 patients with MS and 128 matched healthy controls were enrolled, genomic DNA was isolated and genotyping was performed using mismatched PCR-RFLP. The results were finally analyzed using SPSS. Our results showed significant difference in allelic frequency of SNP rs180515 among cases and controls ($P = 0.004$). For this variation, AA genotype was shown to have protective effect ($P = 0.016$ and OR = 0.6), while GG genotype was a susceptible genotype to MS ($P = 0.04$ and OR = 2.2). Allelic frequency of SNP rs9282641 also showed significant difference between cases and controls ($P = 0.006$). For this SNP, AG genotype had predisposing effect ($P = 0.04$, OR = 2.3), and GG genotype showed protective ($P = 0.01$, OR = 0.411). We successfully replicated the association of two novel SNPs introduced by a GWAS study, and MS in the Iranian population. This result can open ways for better understanding the mechanisms involved in MS.</p>	2017

Let's now scroll back up to the top of this page (also known as Evidence page) and click on the "Multiple sclerosis" link.

Evidence for CD86 in multiple sclerosis



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

Receptor involved in the costimulatory signal essential for T-lymphocyte proliferation and interleukin-2 production, by binding CD28 or CTLA-4. May play a critical role in the early events of T-cell a...

Click on the disease name to get to the disease profile page

[multiple sclerosis](#)
 Synonyms: MS (Multiple Sclerosis), MS, Multiple Sclerosis, Acute Fulminating, Disseminated Sclerosis, MULTIPLE...

By clicking on the disease name, you will get directed to the disease profile page, which contain annotations for the disease (e.g. MS), such as similar diseases (based on common targets), phenotypes, drugs in clinical trials, bibliography and the classification (ontology graph) of the disease.

Open Targets Platform

About Help API Downloads Blog

Search for a target or disease

multiple sclerosis

View associated targets

Synonyms: MS (Multiple Sclerosis) MS Multiple Sclerosis, Acute Fulminating Disseminated Sclerosis MULTIPLE SCLEROSIS ACUTE FULMINATING Sclerosis, Disseminated Sclerosis, Multiple

Similar diseases (based on targets in common)

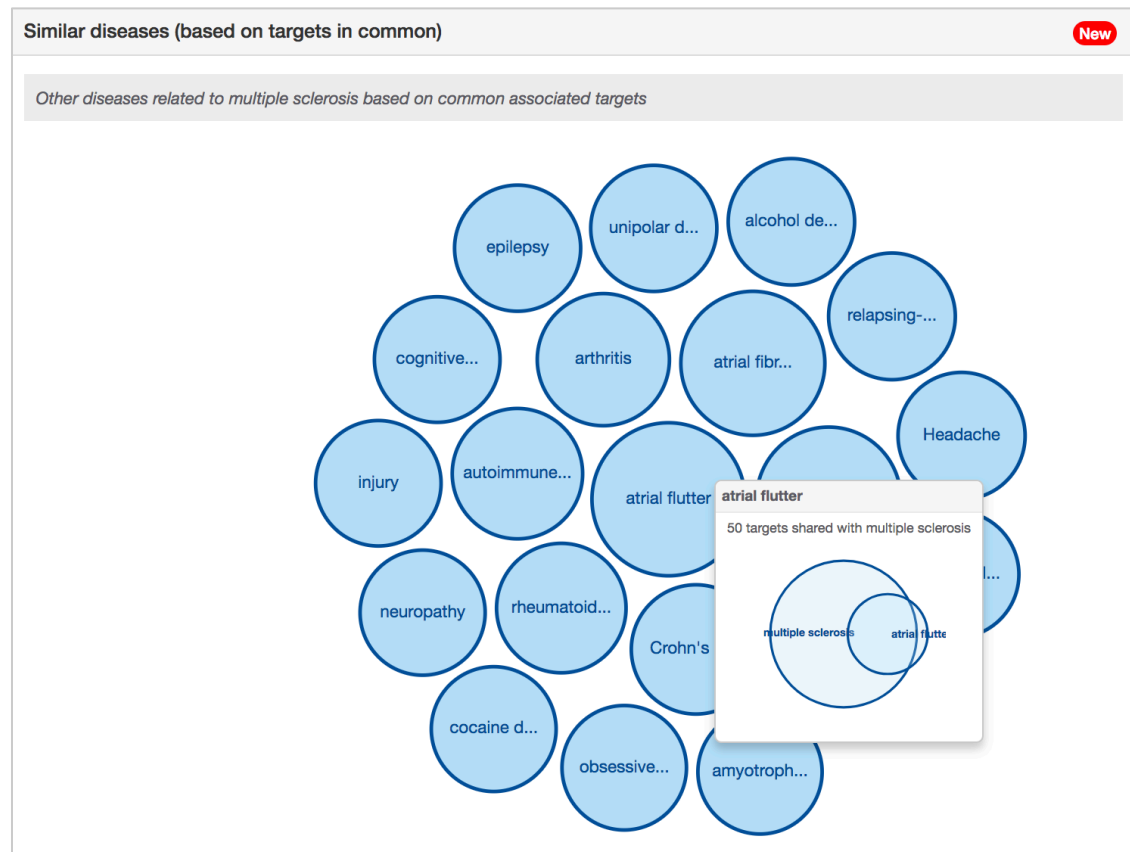
Phenotypes

Drugs

Bibliography

Classification

Let's click on the tab 'Similar diseases (based on targets in common)':

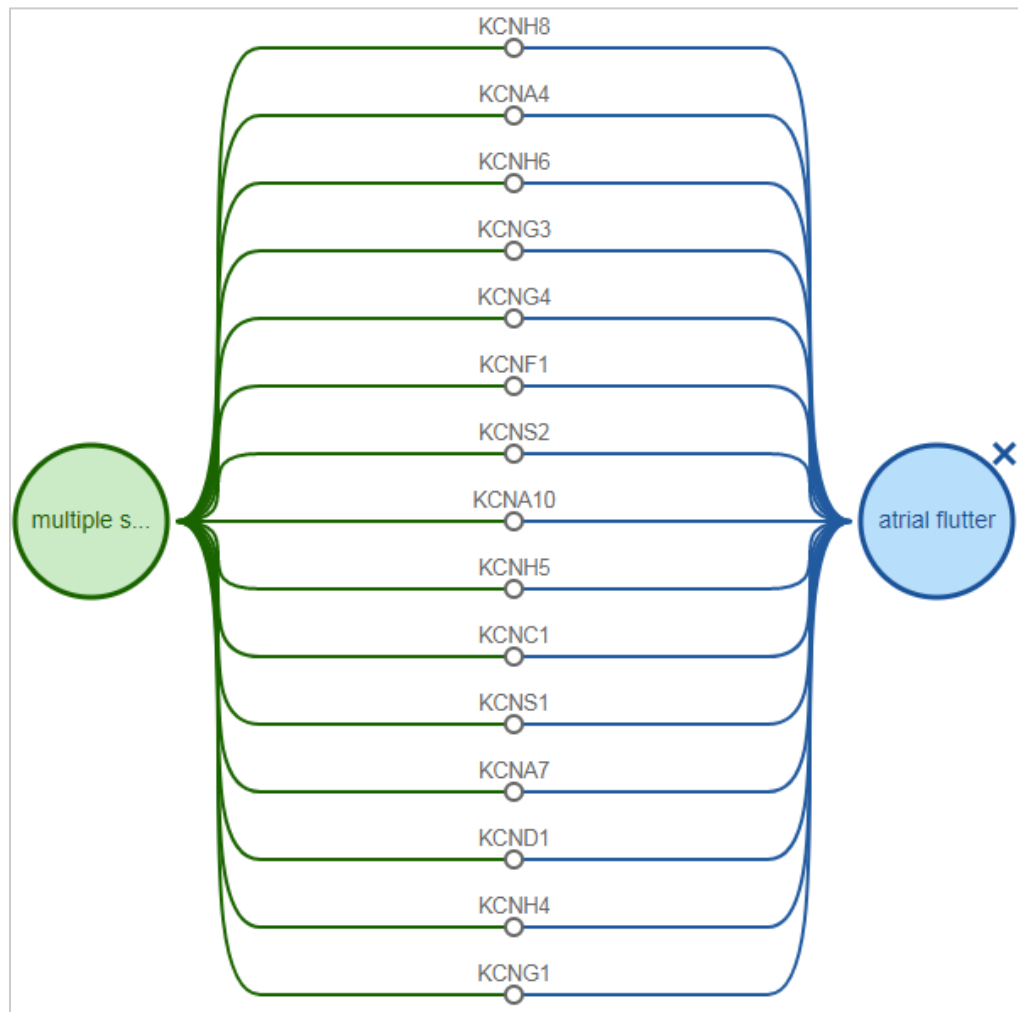


For each pair of diseases, we compute the overlap of shared targets against the total number of connections to both targets, correcting each pair by the significance and the specificity of these connections.

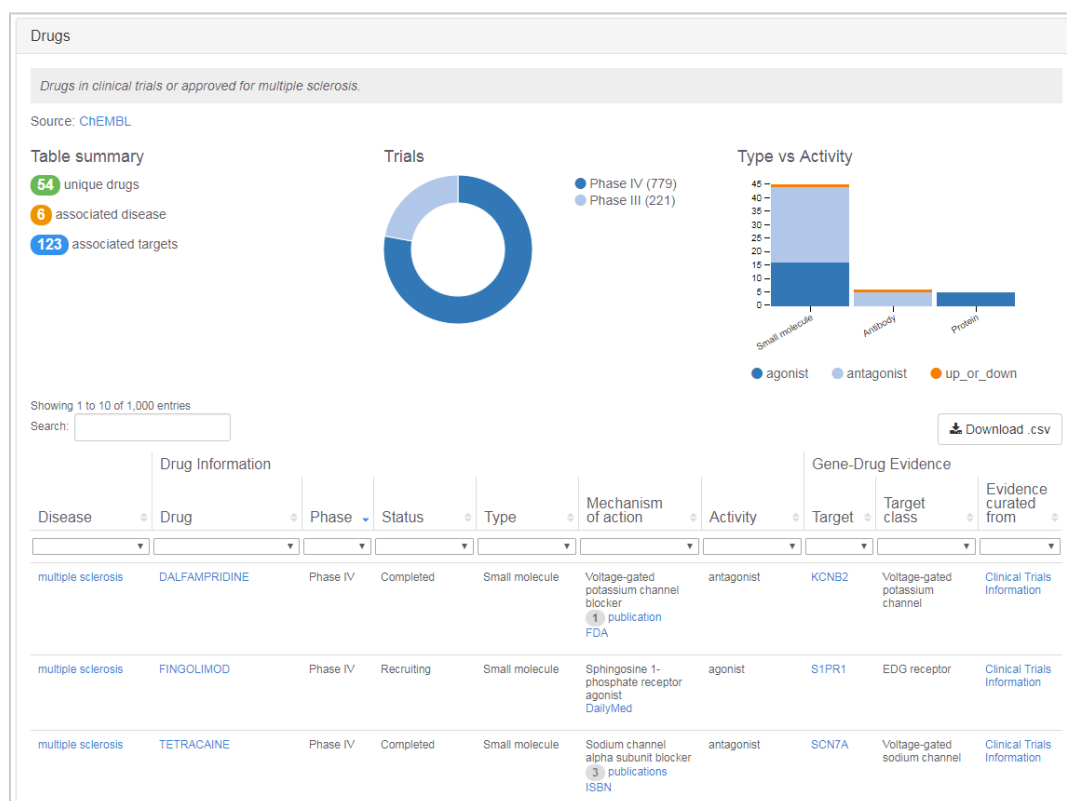
This procedure will consider targets that are specifically linked to fewer diseases more relevant than targets that are commonly linked to many types of diseases.

You will see that the target profile page has a similar visualisation under the tab “Similar targets (based on diseases in common)”. There we will compute a closer distance between two targets sharing a rare disease than two targets sharing diseases that are highly connected to many genes, such as cancer.

You can click on any of the bubbles to get details on the targets in common between any two diseases and the evidence used for the associations (conversely for the diseases in common among any two targets that you can see in the target profile page):



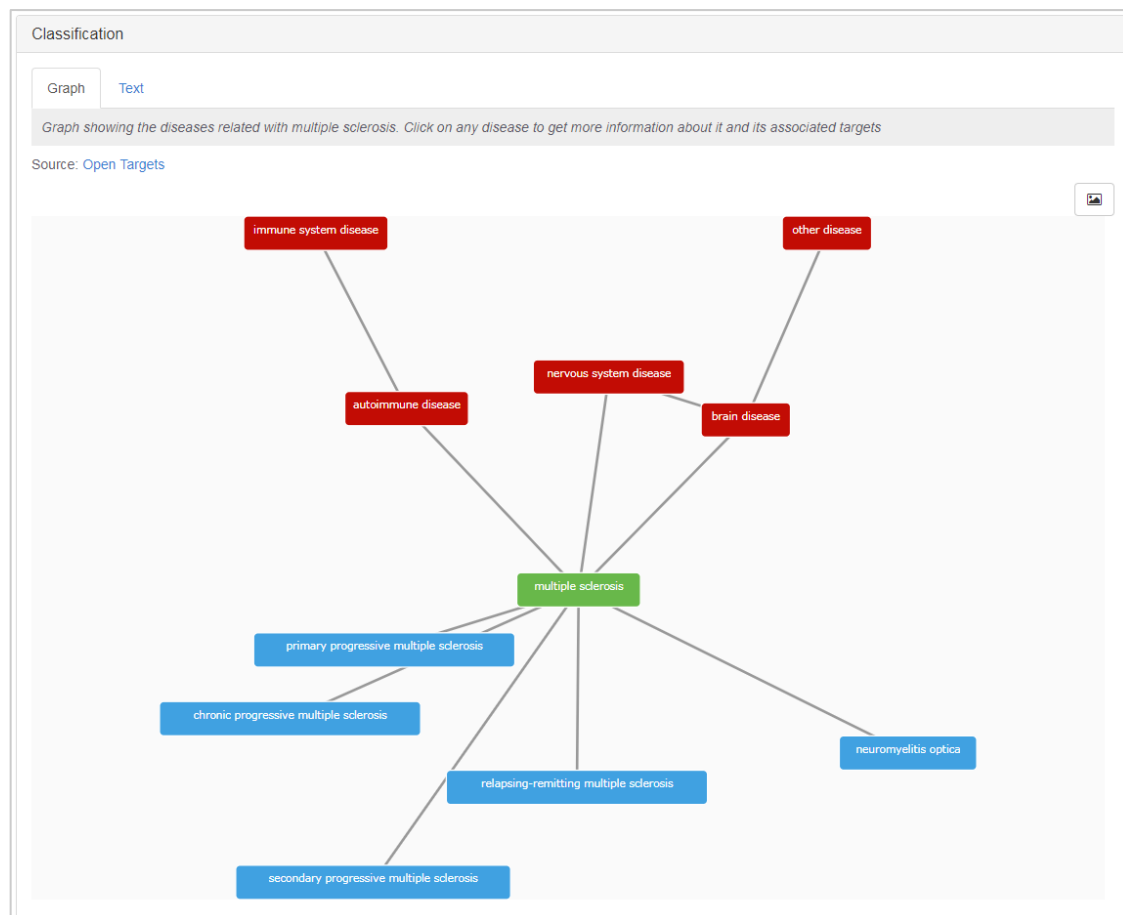
Let's now expand the tab 'Phenotype, then 'Drugs'.



In the August 2018 release, we have more than 50 unique drugs modulating >120 targets and in phase III and phase IV of clinical trials. These clinical trials are carried out with patients suffering from MS or closely related diseases in the ontology of multiple sclerosis, i.e. children terms (e.g. neuromyelitis optica). These drugs will (may) be targeting different proteins.

In addition to the summary visualisation, you will also see a table with drug information. You can filter (and sort) this table by disease, phase of clinical trial (e.g. III), class of the target (e.g. membrane receptor), and much more. You can also download the table in CSV (comma separated value), which can be opened up in Excel.

Still in the disease profile page, scroll down to view the disease ontology (disease concepts and relationships) under the 'Classification' tab. This comes from the EFO (Experimental Factor Ontology), an ontology developed and maintained by EMBL-EBI.



Multiple sclerosis is represented in green. Red nodes correspond to parental terms, whereas the children terms of multiple sclerosis are shown in blue (e.g. chronic progressive multiple sclerosis). Click on any of the disease names to get a pop-up box with the (first) 10 targets associated with any disease:

relapsing-remitting multiple sclerosis

relapsing-remitting multiple sclerosis

EFO code	EFO_0003929
787 genes associated (Showing the first 10)	
CD52	See Evidence
TRPV1	See Evidence
HMGCR	See Evidence
S1PR1	See Evidence

You can then click on any of the options (e.g. “EFO_0003929”, “CD52”, “See evidence”) to explore these specific pages.

By relying on disease ontology and its subtype relationships, we can derive new associations that do not have direct evidence. For instance, IBD is an autoimmune disease that will have direct evidence for its association with its targets. We can propagate this direct evidence up to higher terms in the ontology of IBD and use this evidence (now indirect) to associate target X with autoimmune disease (a parent term of IBD). This procedure can allow us to find common targets across groups of related diseases (e.g. Ulcerative Colitis, Crohn's disease and IBD) even when direct evidence is not available.

For more on this, check our blog post:

<https://blog.opentargets.org/direct-versus-indirect-evidence-should-you-care/>

Now, hit the button “Back” in your browser to go back to the evidence page for CD86 in multiple sclerosis. Then click on the target name, i.e. CD86.



You will land on the target profile page:

[About](#) [Help](#) [API](#) [Downloads](#) [Blog](#)

CD86

CD86 molecule | [View associated diseases](#)

Receptor involved in the costimulatory signal essential for T-lymphocyte proliferation and interleukin-2 production, by binding CD28 or CTLA-4. May play a critical role in the early events of T-cell activation and costimulation of naive T-cells, such as deciding between immunity and anergy that is made by T-cells within 24 hours after activation. Isoform 2 interferes with the formation of CD86 clusters, and thus acts as a negative regulator of T-cell activation.
(information provided by UniProt)

Synonyms: [B7.2](#) [B7-2](#) [CD28LG2](#) [B70](#) [FUN-1](#) [Activation B7-2 antigen](#) [T-lymphocyte activation antigen CD86](#) [CTLA-4 counter-receptor B7.2](#) [B-lymphocyte antigen B7-2](#) [BU63](#)

Drugs

Protein Information

Pathways

Similar targets (based on diseases in common)

Variants, isoforms and genomic context

Protein interactions

RNA and protein baseline expression

Mouse phenotypes

Protein Structure

Gene Ontology

Gene tree

Bibliography

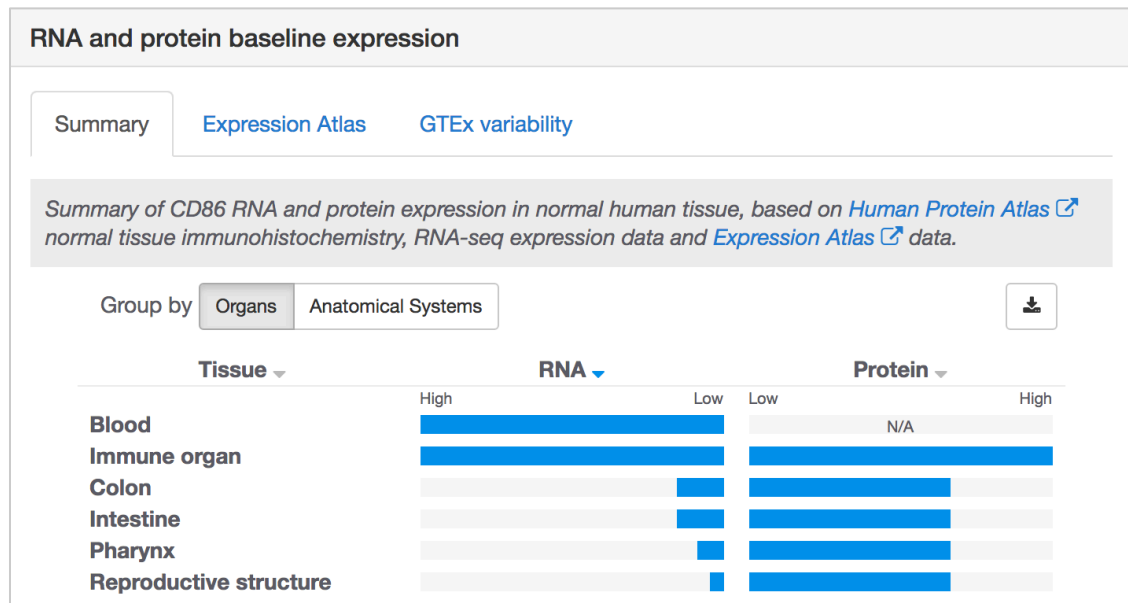
Cancer hallmarks

Cancer biomarkers

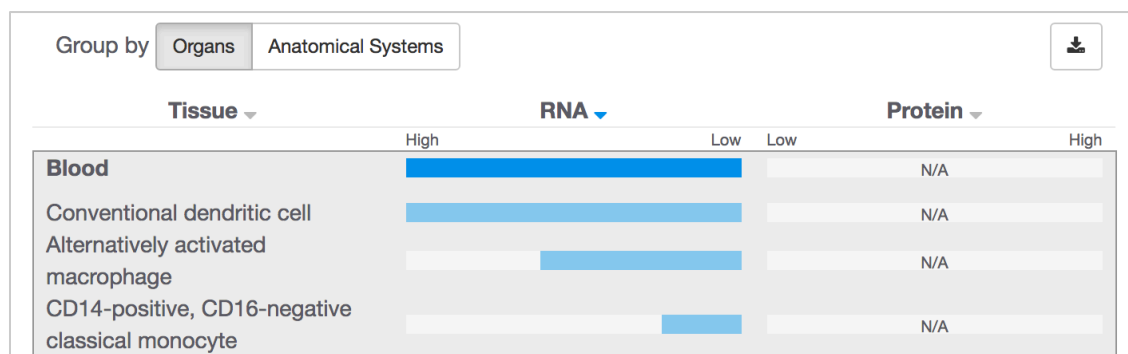
In this page, you can find specific information and annotations for the target, e.g. RNA and protein baseline expression levels, protein structure, gene ontology terms, and plenty more.

Let's expand the 'RNA and protein baseline expression' to find out in which organs or anatomical systems CD86 is highly expressed.

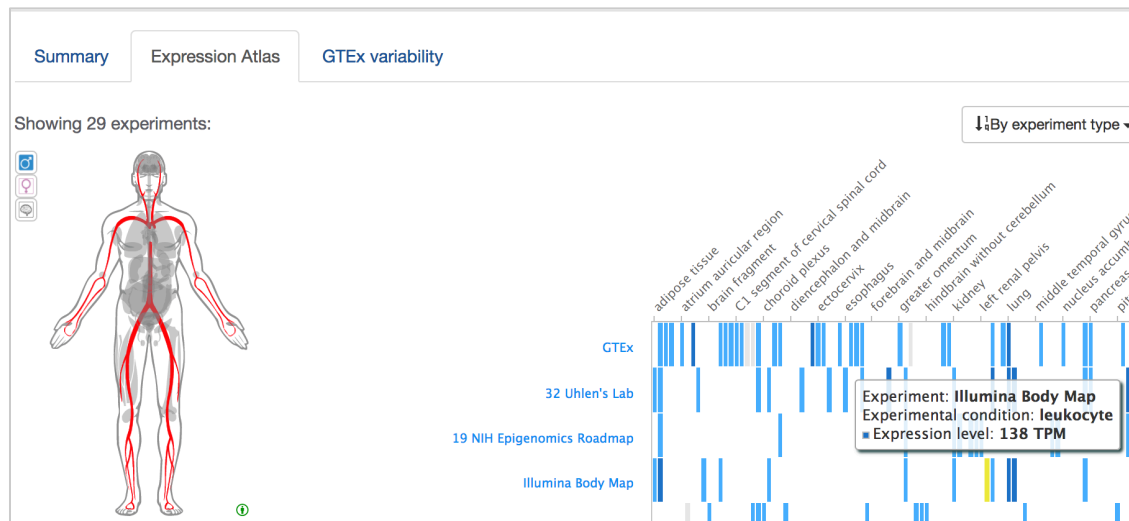
You will find three tabs in there: "Summary", "Expression Atlas" (data from several projects including the Illumina Body Map) and "GTEx variability".



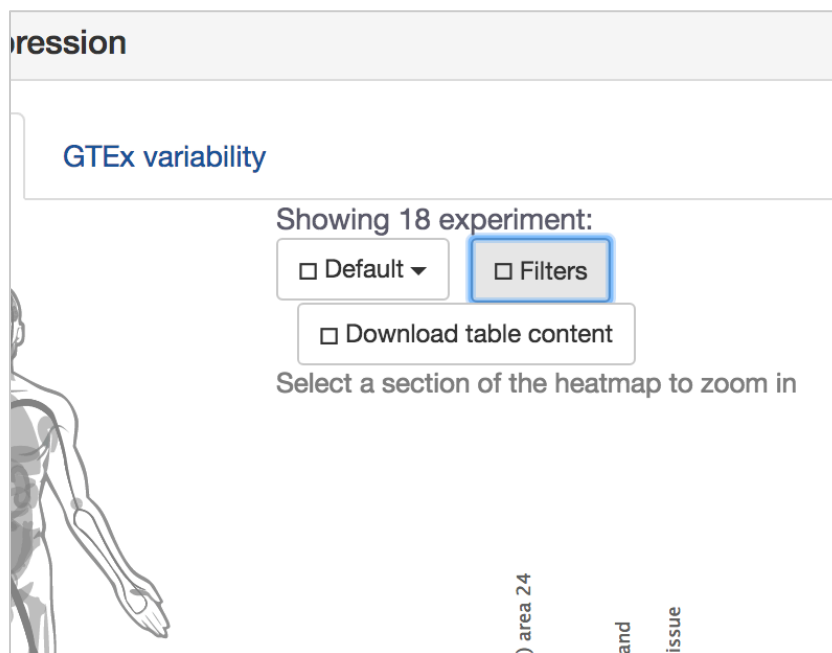
In the Summary tab, you can compare the mRNA and protein expression side by side and learn that the transcript is highly expression at the mRNA level in blood, whereas the protein is highly expressed in the immune organ. You can click on the tissue names to get further granularity such as different type cells in the blood:



Let's now check the 'Expression Atlas' tab for expression data coming from different projects and methodologies (RNASeq and microarray genotyping):



You can zoom in and out and/or apply filters such as 'Expression Value' (below cutoff, low, medium or high) and Anatomical Systems:



Organs

Anatomical Systems

All

High

Medium

Low

Below cutoff

None

☒ Bladder organ ▼

☐ Blood

☒ Brain ▼

☒ Breast


☒ Musculature ▼

☐ Nerve ▼

☒ Oral gland ▼


☒ Pancreas

Let's now view other diseases associated with CD86, apart from MS.


Open Targets Platform

CD86

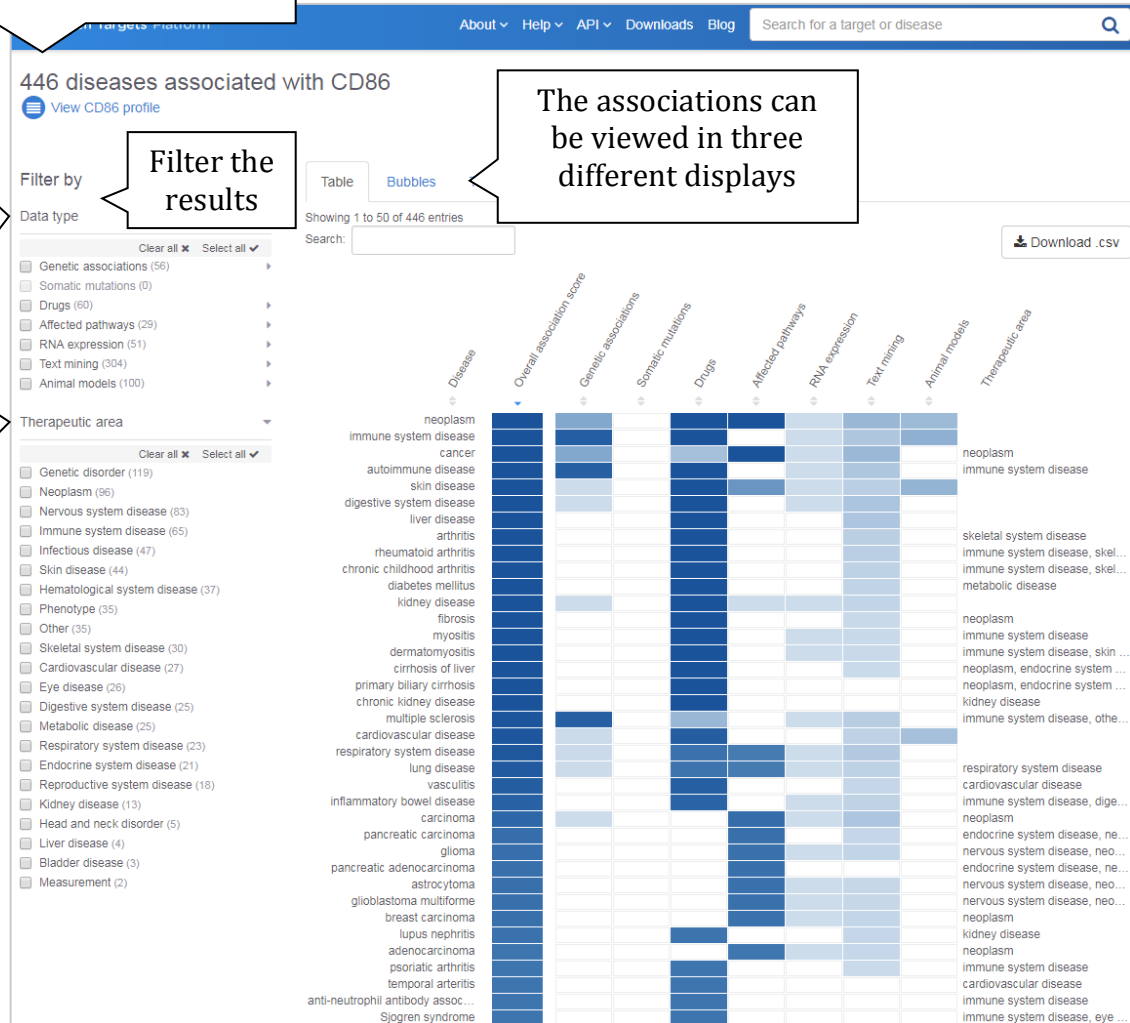
CD86 molecule


View associated diseases

Receptor involved in the costimulatory signal essential for T-cell events of T-cell activation and costimulation of naive T-cells, interferes with the formation of CD86 clusters, and thus acts

Click on 'View associated diseases' to land on a page like this:

Number of diseases associated with CD86



The associations can be viewed in three different displays

Filter the results

Data types (e.g. Drugs)

Therapeutic areas (e.g. Neoplasm)

There are three different displays that can be used to view the diseases associated with any given target:

- Table view

In this view, we list all diseases associated with a target, ordered by the association score, which is colour coded. When there is no evidence to support the association, the cells in this table are coloured in white (score of zero). You can show the 10 first entries and get the pagination for the remaining entries.

This table can be exported in CSV format (look for the download .csv button).

Tip: The different shades of blue in the table convey the strength of the association based on the available evidence (strongest association is represented in dark blue). The score varies from 0 to 1. Hover over the

cells in the table to view the numbers. Alternatively, you can select the cells in the table so that you can view the numerical values.

- Bubble view

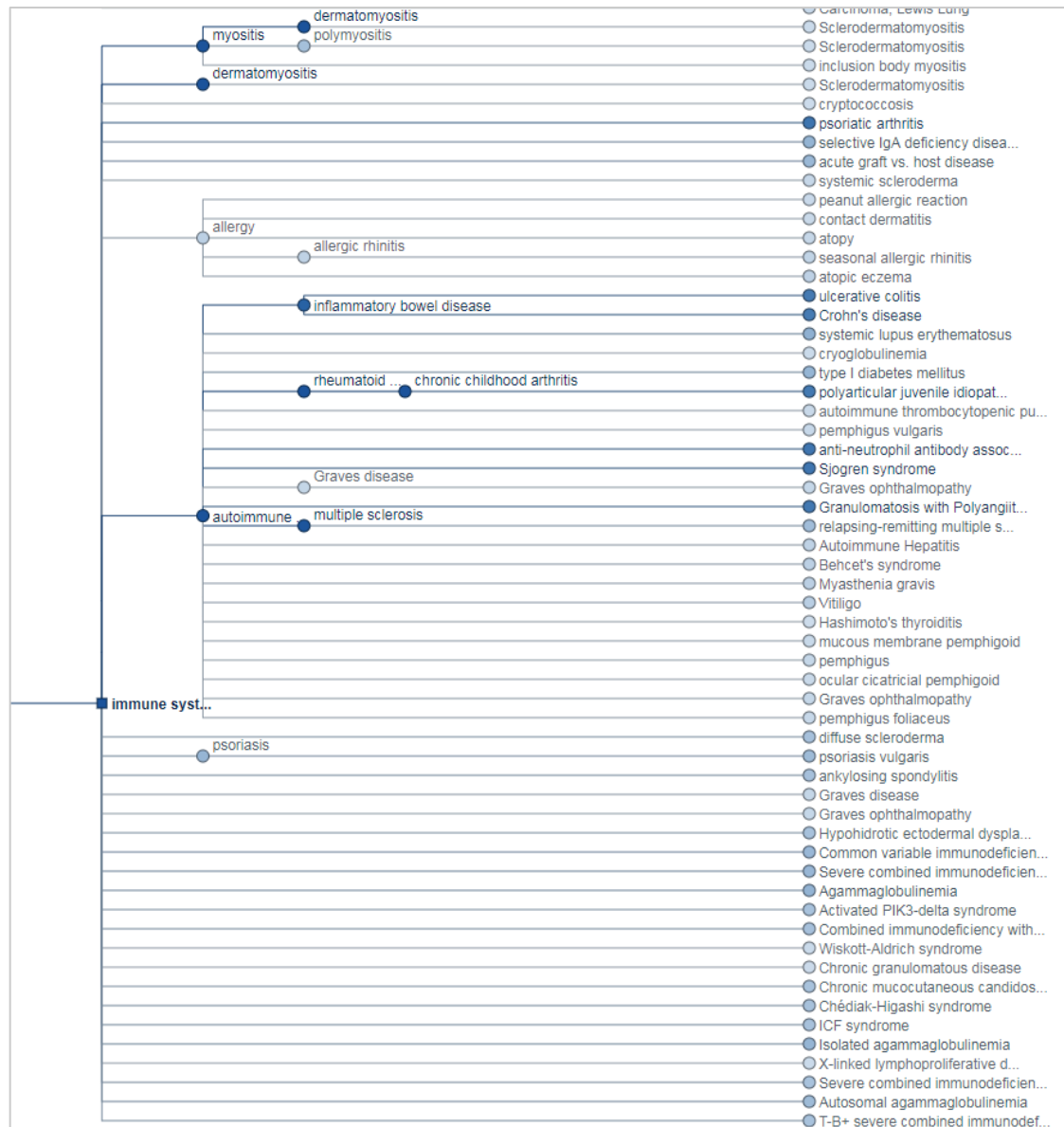
In this view, we group diseases into 'bubbles' based on the disease ontology. Large bubbles correspond to a therapeutic area and consist of smaller bubbles representing diseases within this area. A disease can belong to several therapeutic areas and therefore can appear within more than one large bubble. The strength of the association between the target and a disease is represented by the size of the bubble and the shade of its blue colour; the larger the bubble and the darker the blue, the stronger the association.



- Tree view

In the Tree view, you can visualise the evidence across the therapeutic areas in a tree format that represents the classification of diseases by subtypes. Therapeutic areas have a square symbol (e.g. Genetic

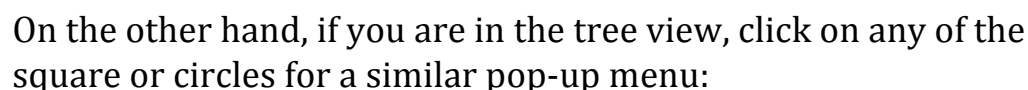
disorders), while the diseases (e.g. multiple sclerosis) are represented as circles. The squares and circles are colour coded in blue, and the darker the blue, the stronger the association:

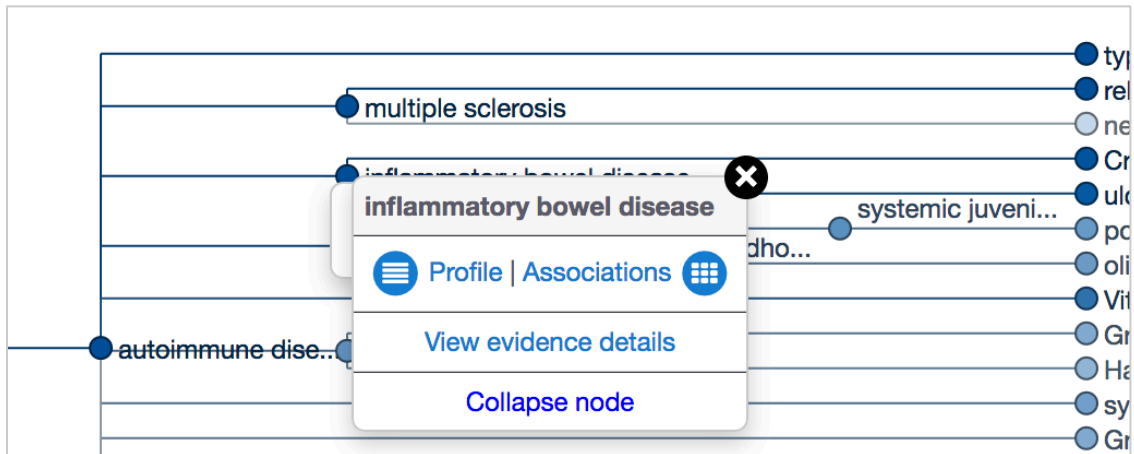


For all these three different views, you have the option to filter the data according to Data type or Therapeutic area. You can, for example, discover other diseases of the “Nervous system” associated with this target.

In the 18.08 release of the Platform, there are 83 of them. These are some examples: ‘Hereditary spastic paraplegia’ and ‘amyotrophic lateral sclerosis’ (ALS). Can you find these two diseases in the table?

C) “View evidence details” to see the underlying evidence for the associations





End of the walkthrough

HANDS-ON EXERCISES

Exercise 1: Vedolizumab and its targets

BACKGROUND

The antibody vedolizumab is in clinical trials to assess its effectiveness and safety as a treatment of adult patients suffering from ulcerative colitis (UC) and Crohn's disease (CD). This has been carried out by one of our partners, Takeda Pharmaceuticals.

QUESTIONS

a) You can search for vedolizumab (and any other drug) using the Open Targets Platform and find out how many targets have associations involving this drug. Which targets do you find in the Platform when searching for that antibody?

b) Let's now focus on the target, returned in the previous question, that has the largest number of diseases associated with. This should be ITGA4. Which data sources support the association of this target with CD? Note the difference between data sources and data types. You will need to expand the Data type options (e.g. Genetic associations, Affected pathways) to see the individual data sources.

c) In addition to vedolizumab, which other drugs targeting this same gene, whether for the treatment of CD or any other diseases, are listed in Open Targets? Explore the different filtering options in the drug information table and feel free to click on to clinicaltrials.gov links in the table for more details on the original trial study.

d) Now click on the Vedolizumab link in the same Drug information table. What is the mechanism of action of this drug? Are there any adverse effects to this drug according to the FDA? When was this drug first approved? In addition to Crohn's disease and ulcerative colitis, can you name other diseases where this drug is associated?

e) Can you name a few targets that are similar ITGA4, based on the set of common diseases between them? Now look for the ITGB1 target, still in the same visualisation. Click on the ITGB1 circle to see the diseases that are in common between ITGB1 and ITGA4. You should see that multiple sclerosis (MS) is one of these diseases in common. Can you find out which data types have provided us with evidence for the associations between these two targets and MS?

e) In which tissues ITGA4 has the highest level of baseline mRNA expression? You may want to click on the tissue name (or the blue bar in the histogram) to see the different cell types we have summarised the mRNA expression data for.

f) Have a look at the mouse phenotypes tab now to see the observed phenotypes in mice that have had ITGA4 knocked out. If you were working in the safety assessment of this target, based on the knockout mouse phenotype data, what would be your recommendation on modulating the function of ITGA4? Would it be safe to target this protein? Which impact it might have in humans?

g) Which are the latest papers on this target that we list in our Platform? Explore the different filters available under Bibliography (e.g. Concepts, Genes, Diseases, etc). Choose one of them and click on “Show abstract” to see the annotations (text highlighted) available. You may want explore the similar articles available for any given paper. Hint: our bibliography is the result of a NLP (natural learning processing) of more than 25 million abstracts in PubMed. This analysis provides semantic relationships between genes, diseases, drugs, biological concepts and other entities. Check more details on this tool, called LINK, in our [blog](#).

Exercise 2: Advancing research in the field of IBD

BACKGROUND

More than five million people worldwide live with IBD (inflammatory bowel disease), whose symptoms can be unpredictable. While the causes of IBD are unknown, several hypotheses have been suggested so far including genetic predisposition, environmental triggers, immune system, and chronic and aberrant inflammation. Go to the Open Targets Platform to answer the following:

QUESTIONS

a) How many targets associated with IBD, involved in the interleukin-4 and 13 signaling pathway (within immune system), with evidence supported by common variants in the GWAS catalog?

b) TYK2 is one of the targets associated with IBD retrieved from the previous question. Let's now look closely at the evidence behind this association and answer the following:

- How many variants are listed as being associated with IBD?
- Can you determine if these are direct or indirect evidence?
- What is the source of this information? Can you retrieve the exact scientific publication(s) where this association was curated from?

c) Can you see these variants displayed on the genome, in a browser like view? Have a look at the different colours used to represent the variants. Can you find out what variants highlighted in red mean? Are there other diseases, in addition to IBD, Crohn's disease and ulcerative colitis, associated with these variants highlighted in red? Check how significant the associations are (look for the p-value).

c) How many drugs targeting TYK2 are in clinical trials with patients suffering from IBD? Why are Crohn's disease and ulcerative colitis listed in this table?

Exercise 3: Filtering Alzheimer's disease associations based on a list of targets

BACKGROUND

A drug discovery scientist at Alzheimer's Research UK has a list of eight targets that seem to be associated with Alzheimer's disease (AD) based on literature reviews. These are HFE, PSEN1, PRO1557, APOE, ADRB2, PSEN2, CPAMD5, BACE1. You can download the list from:

https://github.com/deniseOme/training/blob/master/target_list_file.txt

QUESTIONS

a) Once you downloaded the list and saved it as .txt you should upload it in page showing the targets associated with Alzheimer's. Look for the filter called "Your target list", click on "Choose file" and select the

file you have downloaded. Which of those eight targets have higher levels of mRNA expression in the brain than in any other tissue? Note: this is known in the Open Targets Platform as 'RNA tissue specificity'.

b) Are any of these targets expressed specifically in the brain classified as secreted protein? Note: you can filter targets based on the Target class filter.

c) Let's now explore this target in more detail and answer the following:

- Does this protein have antigenic sequences? Which amino acids do these sequences correspond to?
- Which pathways is this protein involved with?
- Are there any mutations in this gene that have been associated with AD?

Exercise 4: LRRK2 in Parkinson's disease

BACKGROUND

The LRRK2 gene encodes a protein with five putative functional domains: an N-terminal leucine-rich repeat (LRR) domain, a Roc (Ras of complex protein) domain that shares sequence homology to the Ras-related GTPase superfamily, a COR (C-terminal of Roc) domain, a mitogen-activated protein kinase kinase kinase (MAPKKK) domain, and a C-terminal WD40 repeat domain. A genetic variant in this gene is one of the most common causes of inherited Parkinson disease (Gandhi et al., 2008).

QUESTIONS

a) How long is the protein encoded by this gene/target? Can you find the protein domains listed above?

b) No drug is currently available to target LRRK2. There may be other compounds such as chemical probes? Can you use the Open Targets Platform to see which chemical probes, if any, are available that could be used to modulate the function of this protein?

c) Can you list some of the proteins that interact with LRRK2? Can you download this image?

d) Let's now have a look at the diseases associated with this target. Can you download a table containing all diseases from the nervous system where there is evidence for the associations from genetics and animal models?

EXTRA HANDS-ON EXERCISES

Exercise E1: Assessing the specificity of targets for IBD

BACKGROUND

A biologist working on translation medicine at EMBL has a list of 26 genes linked to IBD. However they don't know how specific this list is and if whether or not some of these genes could be therapeutic targets for other diseases of the digestive system.

What is the best way to find this out using the Open Targets Platform?

The list of genes can be found on:

https://github.com/deniseOme/training/blob/master/list_IBD_batch_search.csv

Exercise E2: Getting all disease associations and scores for three targets at once

BACKGROUND

You have used the website www.targetvalidation.org to search for target-disease associations on a one-by-one and the batch search to find the diseases, pathways, GO terms represented in your list of targets. Now you have three genes you would like to find all diseases associated with them based on direct evidence.

ENSG00000141736

ENSG00000141510

ENSG00000132356

QUESTIONS

a) Can you retrieve all diseases associated with those three Ensembl gene IDs? What is the total number of diseases retrieved for these three genes?

b) Can you now filter these associations to contain evidence from the data type "Genetic associations"? How many diseases are associated with these three targets based on 'dataype=genetic_association' only?

c) Can you download the above list in TAB format?

QUICK GUIDE TO DATABASES

Here is a list of databases and projects that may be useful for you. Some of them are used as [data sources](#) for gene-disease associations available through our Open Targets Platform.

PROTEINS

UniProtKB – The “Protein knowledgebase” is a comprehensive set of protein sequences. It is divided into two parts: TrEMBL and Swiss-Prot. The later is manually annotated and reviewed, therefore provides a set of protein sequences of high quality.

<http://www.uniprot.org/>

GENE NOMENCLATURE COMMITTEES

HGNC – The HUGO Gene Nomenclature Committee assigns unique names and symbols to every single human gene, whether they are coding or not. These gene names and symbols are the official ones for human genes.

<http://www.genenames.org/>

MGI – The HGNC counterpart for naming mouse genes and symbols.

<http://www.informatics.jax.org/>

GENETIC VARIANTS and SOMATIC MUTATIONS

GWAS Catalog– The catalog of Genome Wide Association Studies (GWAS) provides genetic variants (e.g. SNPs) that are associated with a disease.

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

EVA – The European Variation Archive (EVA) provides genetic variants and somatic mutations (associated with cancer).

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

Cancer Gene Census – A catalogue of genes for which mutations have been causally implicated in cancer. The Catalogue of Somatic Mutations in Cancer (COSMIC) at the Wellcome Sanger Institute provides us with the set of genes associated with specific cancers in the Cancer Gene Census, in addition to other cancers associated with that gene in the COSMIC database.

www.cancer.sanger.ac.uk/census/

COSMIC is also the database that provides us with the cancer hallmarks:

<https://cosmic-blog.sanger.ac.uk/hallmarks-cancer/>

IntOgen - It provides evidence of somatic mutations (driver mutations), genes and pathways involved in cancer biology from 6,792 samples across 28 cancer types.

<https://www.intogen.org/search>

Gene2Phenotype - The data in Gene2Phenotype (G2P) provides evidence of genetic variants that are manually curated from the literature by consultant clinical geneticists in the UK. This is provided by DECIPHER, a database of genomic variants and phenotypes in patients with developmental disorders.

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

Genomics England PanelApp - The Genomics England PanelApp is a knowledgebase that combines crowdsourcing of expertise with curation to provide gene-disease relationships to aid the clinical interpretation of genomes within the 100,000 Genomes Project.

<https://panelapp.extge.co.uk/crowdsourcing/PanelApp/>

PheWAS Catalog

The PheWAS (phenome-wide association studies) resources provide associations between a genetic variant and multiple phenotypes. It contains clinical phenotypes derived from the electronic medical record (EMR)-linked DNA biobank BioVU by the Center for Precision Medicine at the Vanderbilt University Medical Center.

<https://phewascatalog.org/>

DRUGS

ChEMBL - The ChEMBL database at the EMBL-EBI provides evidence from known drugs that can be linked to a disease and a known target.

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

RNA EXPRESSION

Expression Atlas - The Expression Atlas at EMBL-EBI provides information on genes that are differentially expressed between normal and disease samples, or among disease samples from different studies. In addition to differential expression, they provide baseline expression information for each gene.

<https://www.ebi.ac.uk/gxa/home>

AFFECTED PATHWAYS

Reactome - The Reactome database at the EMBL-EBI contains pathway information on biochemical reactions sourced from manual curation. It identifies reaction pathways that are affected by pathogenic mutations.

<http://www.reactome.org/>

SLAPenrich - It's a statistical framework for the identification of significantly mutated pathways, at the sample population level. We include in the Open Targets Platform the data obtained using SLAPenrich on somatic mutations from the The Cancer Genome Atlas across 25 different cancer types and a collection of pathway gene sets from Reactome.

<https://saezlab.github.io/SLAPenrich/>

PROGENy - [PROGENy](#) (Pathway RespOnsive GENes) is a linear regression model that calculates pathway activity based on consensus gene signatures obtained from perturbation experiments. We use PROGENy ([Schubert et al](#)) for the systematic comparison of pathway activities between normal and primary samples from The Cancer Genome Atlas (TCGA). We include in our Open Targets Platform sample-level pathway activities inferred from RNA-seq for 9,250 tumour and 741 normal TCGA samples from 14 tumour types, and compute differential pathway activities between matched normal and tumour samples. We cover the following pathways: EGFR, hypoxia,

JAK.STAT, MAPK, NFkB, PI3K, TGFb, TNFa, Trail, VEGF, and p53. See [Schubert et al \(2018\)](#) for more details.

TEXT MINING

Europe PMC - The Europe PubMed Central at the EMBL-EBI mines the titles, abstracts and full text research articles from both PubMed and PubMed Central to provide evidence of links between targets and diseases.

<http://europepmc.org/>

ANIMAL MODELS

Phenodigm - Phenodigm is an algorithm developed by Damian Smedley at the Wellcome Trust Sanger Institute that use a semantic approach to map between clinical features observed in humans and mouse phenotype annotations. The results are made available on the IMPC portal:

<https://www.mousephenotype.org>