Mining gene and disease associations with Open Targets for improved target identification

Coursebook

Bristol-Myers Squibb
30th May 2017

Denise Carvalho-Silva

Open Targets
Wellcome Genome Campus
Hinxton, United Kingdom

Notes

This course is based on the April 2017 release of our Open Targets Platform.

Some useful links:

1) About the Open Targets Consortium
www.opentargets.org/about

2) About the Open Targets Platform
www.targetvalidation.org/about

3) Our publication
www.bit.ly/OpenTargets

4) Details on the latest Platform release
http://blog.opentargets.org/extra-data-in-our-open-targets-platform/

Feel free to tackle questions relative to your own research instead of following the ones provided in this coursebook.

Questions or Feedback?
support@targetvalidation.org

TABLE OF CONTENTS

OVERVIEW

Open Targets is a public-private initiative to generate evidence on the validity of therapeutic targets based on genome-scale experiments and analysis. We are working to create an R&D framework that applies to a wide range of human diseases, and we want to share this data 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 Trust Sanger Institute (http://www.sanger.ac.uk/) and the European Bioinformatics Institute (http://www.ebi.ac.uk/). In February 2016, a forth institution namely Biogen (https://www.biogen.com/) joined the initiative and the consortium was rebranded to Open Targets in April 2016.

In the process of drug discovery, the *validation* of a target refers to the creation of a specific entity that modulates that target's activity 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 failures than successes. 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 projects of Open Targets consortium
- An introduction to the Open Targets Platform
- Browsing the Platform
- Pointing to alternative ways to access the data

INTRODUCTION TO OPEN TARGETS

Open Targets employs large-scale human genetics and genomics data to change the way drug targets are identified and validated. We have established a set of projects to develop both the data and analytical processes that implicate targets as valid, and the core platform to provide the information to a diverse audience of users.

The core bioinformatics team develops pipelines and a database to integrate existing target data. The core also designed, created and maintains the Open Targets Platform, a public web portal to serve the integrated data and views.

Our experimental projects focus on providing insights in the identification of targets relevant to key therapeutic areas namely:

- Oncology
- Inflammatory bowel diseases (IBD)
- Respiratory disease
- Inflammation and immunity
- Neurodegenerative diseases

We also aim to develop standard epigenome profiles of cell models in use within the pharmaceutical industry and academia and establish a systematic approach for the determination of human biological and disease relevance.

More details can be found in our Projects page.

Retrieving data from Open Targets with our Platform

The Open Targets Platform is a web application that integrates and displays publicly available biological data to foster the discovery and prioritisation of targets for new therapies. We use data sources as diverse as Gene2Phenotype, IntOGen, GWAS, UniProt, ChEMBL, Expression Atlas, Cancer Census, Reactome and EuropePMC as pieces of evidence to support target-disease associations. The associations are scored using objective statistical and computational techniques.

In our latest release (April 2017), the Platform provides information on 31,380 targets; 8,891 diseases; 5.3 million evidence; and 2.6 million target-disease associations.

In addition to the web application, we include the data dumps, a REST API and a Python client.

The Open Targets Platform is aimed at users from both academia and industry, whether they want to browse a target on a gene by gene (or disease by disease) basis, carry out more complex queries using the API, or download all evidence and association objects for downstream analyses.

Synopsis: what can I do with the Open Targets Platform?

- Find out which targets are associated with a disease
- Explore the evidence supporting this target-disease association
- Export a table with the FDA drugs currently in clinical trials
- Discover if there other diseases associated with a given target
- Get the association of a target with diseases from different therapeutic areas
- Find target specific information, such as baseline expression, protein structure, alternatively spliced transcripts, gene trees
- Get disease target specific information, such as a classification based on the ontology of the disease and the drugs mapped to it

Help documentation and support

- Data sources in the Open Targets Platform
- View our FAQs
- Email us

Connect with us

- Open Targets Blog
- Follow us on Twitter
- Check our page on Facebook and LinkedIn

Further reading

Koscielny, G. *et al.* Nucleic Acids Res (2017 Database Issue): http://nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

A breakthrough article from Nucleic Acids Research: http://www.narbreakthrough.com/

OPEN TARGETS PLATFORM: LIVE DEMOS

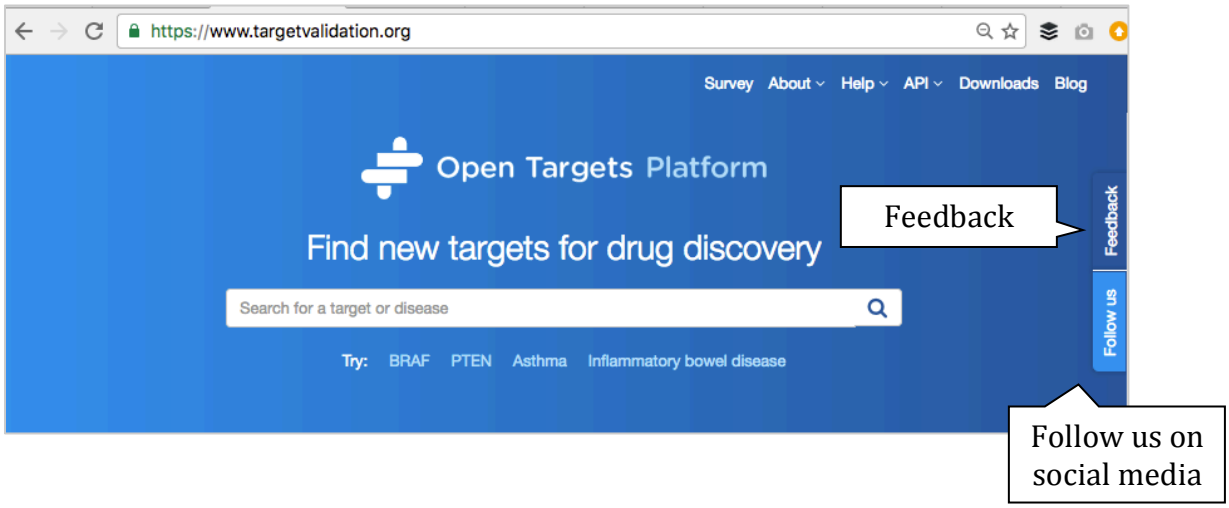
We will guide you through the website using multiple sclerosis, as an example and exploring the *CD86* gene.

The following points will be addressed during the walkthrough:

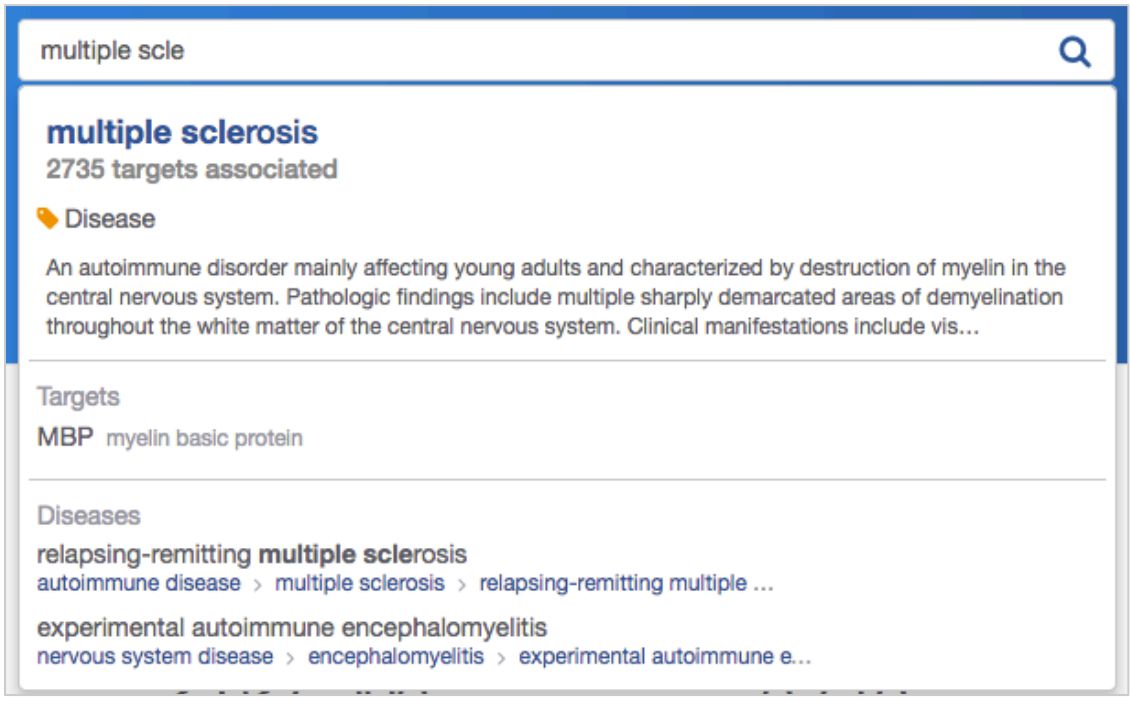
- How to find targets associated with multiple sclerosis
- How to filter down the number of targets based on specific type of evidence
- How to find the data sources used to support the target-disease association
- How to look for other diseases (than multiple sclerosis) associated with a given target
- How to visualise a target in a browser-like view
- How to find out how strong the association between a target and a disease is
- How to find drugs currently in clinical trials
- How to filter **our** associations by **your** list of target genes

Demo 1: Disease centric workflow

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



Select the first (best) hit:



You will see a page like this:

The current release of the Open Targets Platform (April 2017) lists 2778 targets associated with multiple sclerosis.

The data types supporting these results are based on Genetic association, Somatic mutations, Drugs, RNA expression, Text mining, and Animal models. There is no data on affected pathways available for this disease.

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

You can filter the number of associations by 'Data types', 'Pathway types', 'Target class' and by uploading 'Your target list':

A) Data types

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

B) Pathway types

- Signal Transduction

Metabolism

…

C) Target class

Enzyme

Membrane receptor

…

D) Your target list

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

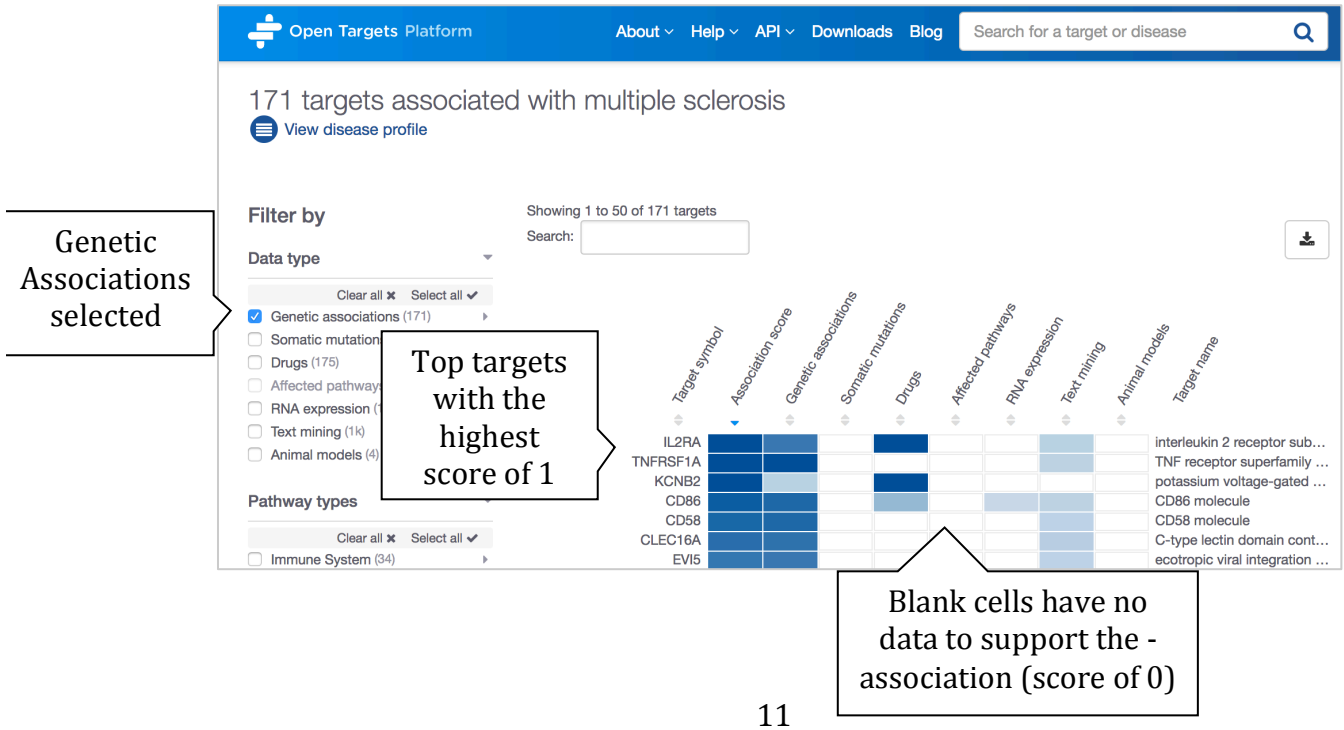
What are **Data types, Pathway types** and **Target class**?

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' are defined by Reactome:
http://www.reactome.org/

whereas the categories within 'Target class' are defined by ChEMBL https://www.ebi.ac.uk/chembl/.

Let's now filter the data to focus on 'Genetic associations' (Data type) only. The number of targets goes down to 171:



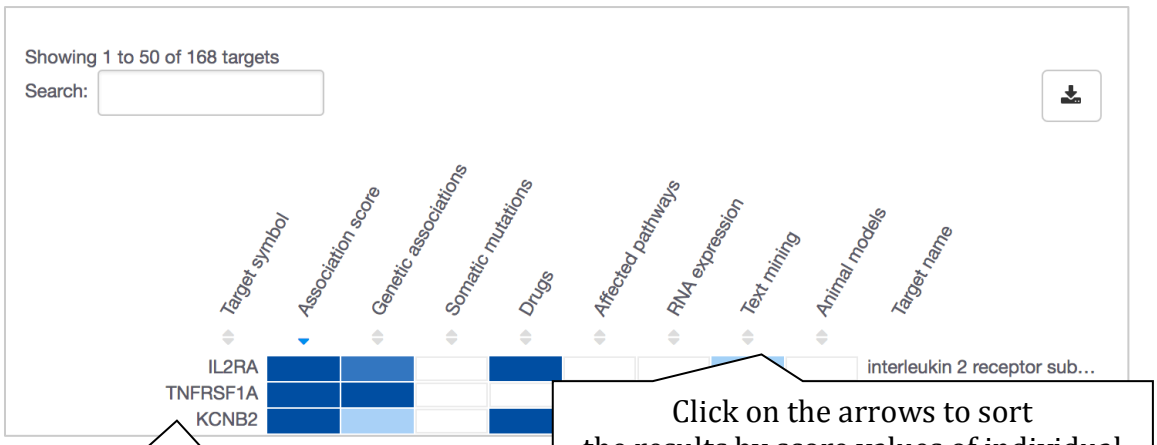
Those are targets associated with multiple sclerosis based on genetic variants only. Genetic variants are SNPs (single nucleotide polymorphisms or mutations, or short indels) from the GWAS catalog, UniProt and EVA (European Variation Archive) sources only.

The table is sorted by default with the best hit on the top of the table i.e. *IL2RA*. The best hit is the target that contains the highest (overall) association score. Different weight is given to different data types when computing the score:

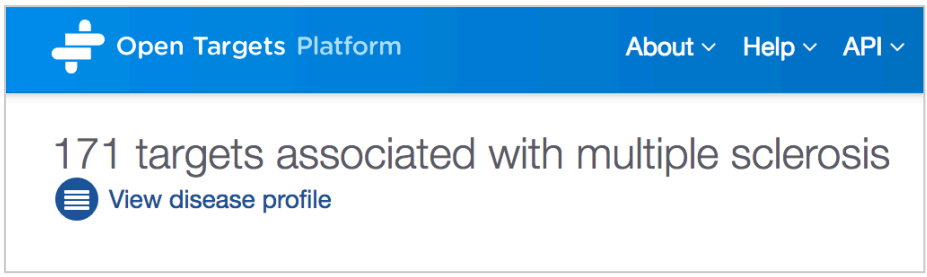
From the highest to the lowest weight, this is the ranking of our data types:

Genetic association = somatic mutations = drugs = pathways
RNA expression
Animal models = Text mining

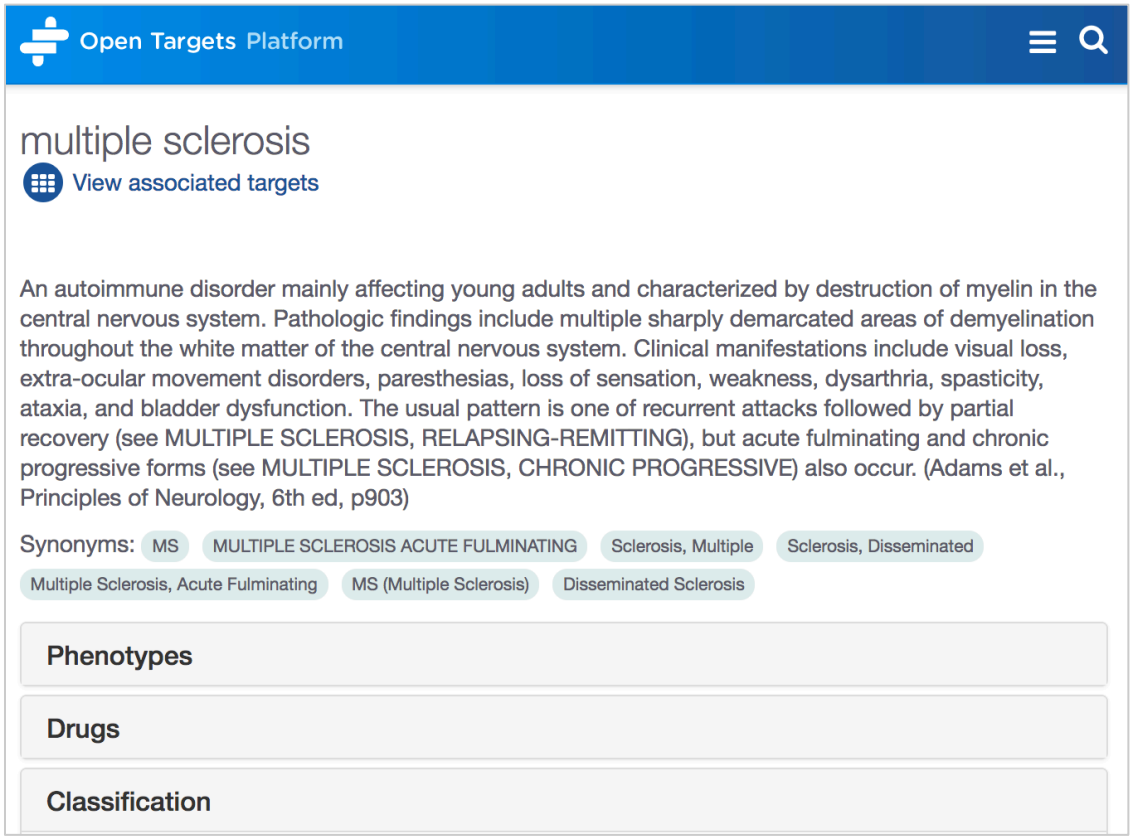
You can sort the table by alphabetical order of the list of targets, or by the association score (either overall or per data type e.g. Genetic associations, Drugs, Text mining, etc). The association score varies from 0 to 1, the closer to 1 the stronger the association. This score is computed for each 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 ('Association score' column in the table below):



To see which drugs are used in the clinical trials with patients with multiple sclerosis (or children terms of this disease), click on the 'View disease profile' below:

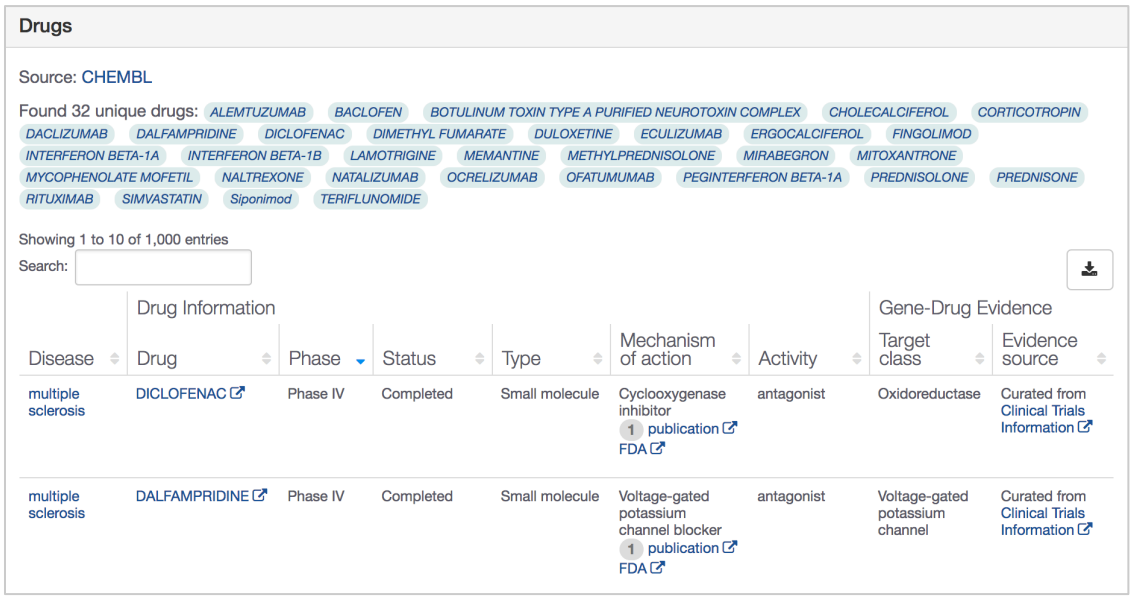


Let's now expand the tab 'Drugs'. You will see a page like this:

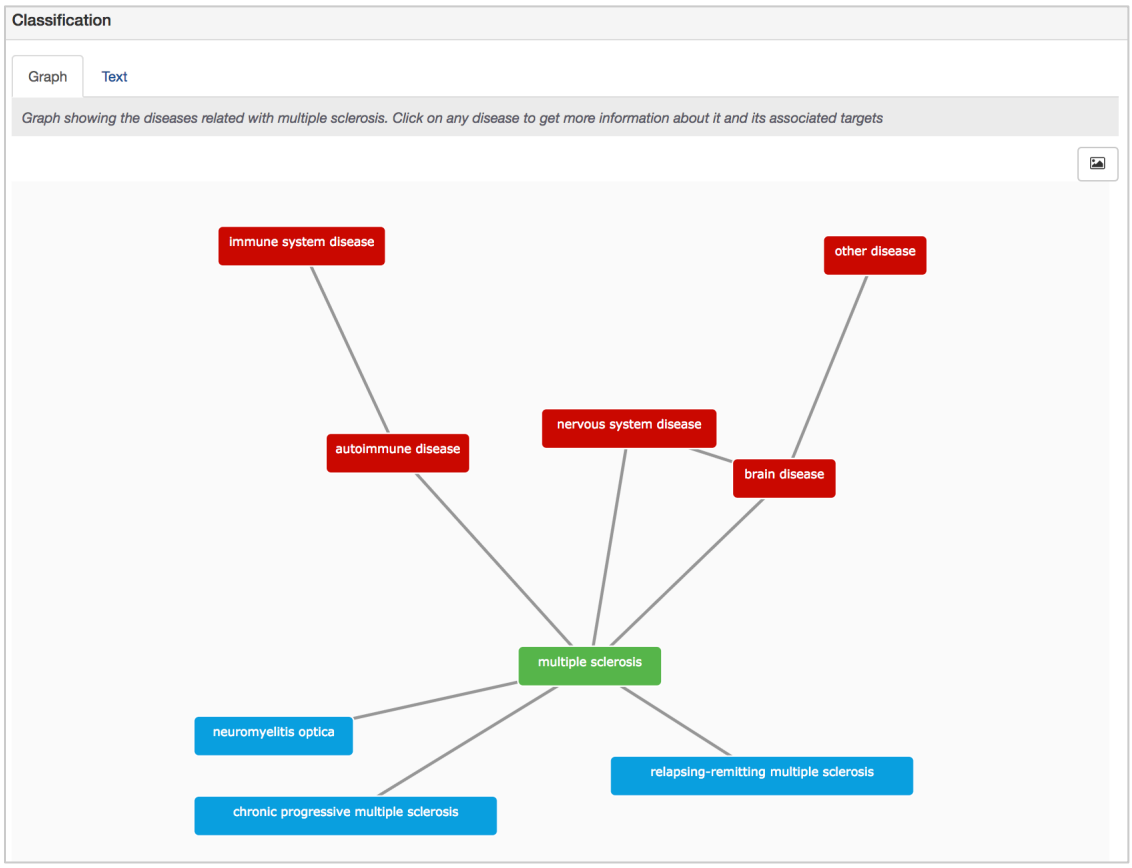


In the April release of the Open Targets Platform, we have 31 unique drugs in different phases of clinical trials with patients suffering from this condition. They will be targeting different proteins.

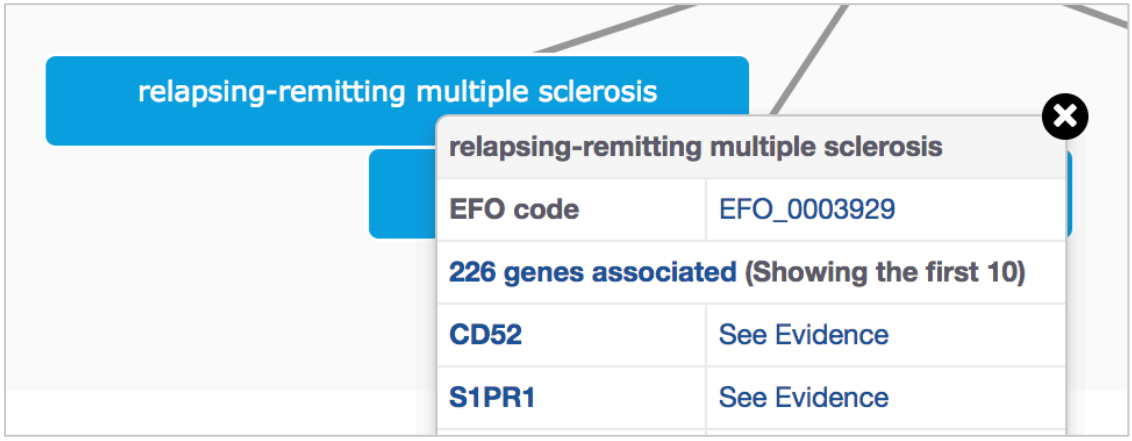
You can filter (and sort) the table by disease, phase of clinical trial (e.g. IV, the most advanced phase), class of the target (e.g. membrane receptor), and much more. You can also download this table in csv (comma separated value):



Next we can scroll down to view the disease ontology (disease relationship) in the 'Classification' tab. The data on the relationship of diseases comes from the EFO (Experimental Factor Ontology) and the diagram is developed by Open Targets.



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. chromic progressive multiple sclerosis). Click on any of disease names to get the targets associated with them:



By using the EFO parent-child relationships, we can derive new associations that do not have direct evidence. For instance, IBD is an autoimmune disease and the direct evidence for targets associated with IBD are propagated to the higher autoimmune level to allow

users to find common targets across groups of related diseases (e.g. Ulcerative Colitis, Crohn's disease and IBD). In EFO, 'asthma' is a 'respiratory system disease' and 'childhood onset asthma' is a subclass of 'asthma'. Both evidence from 'asthma' and 'childhood onset asthma' are propagated to 'respiratory system disease'.

So you will get direct and indirect associations of diseases and genes. For more on this, check our blog post:

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

Demo 2: Evidence supporting association

Let's now explore the evidence used to associate *CD86* and multiple sclerosis. We've chosen this target for follow up based on the availability of information from different data types.

Click on any row corresponding to CD86 in the table below:

land in the evidence page for the association between a gene and a disease:

The data types that support this association are (see the flower and the coloured petals):

Genetic association
Drugs
RNA expression
Text mining

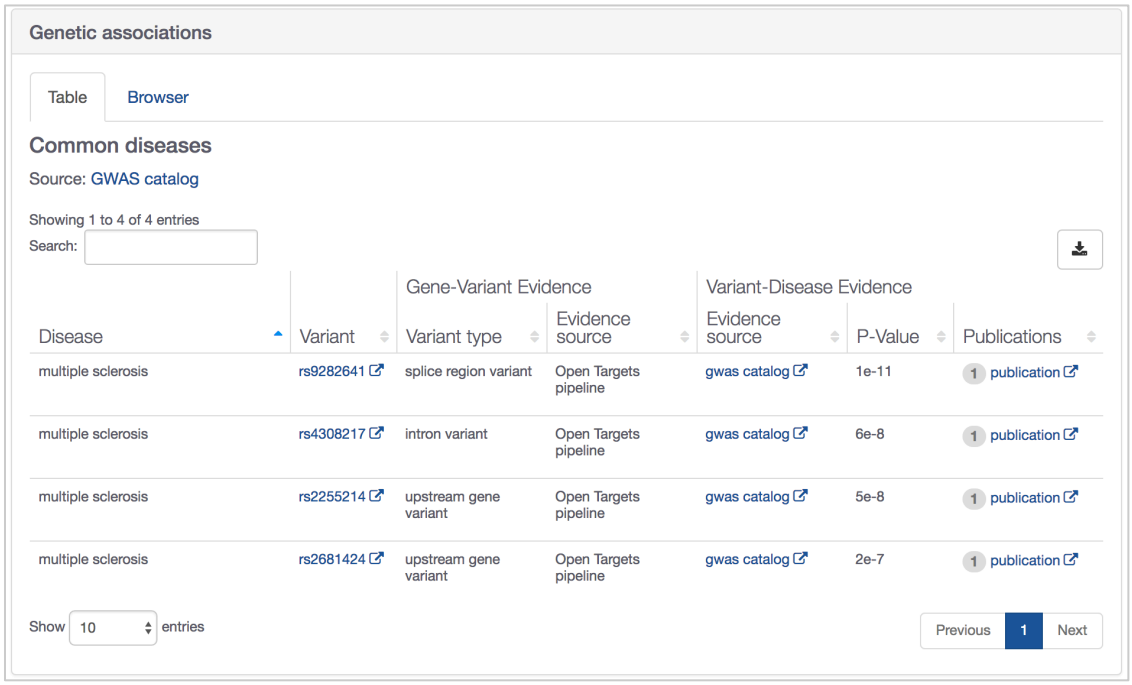
Note: If you wish to suggest data types or resources we could/should incorporate into our Platform, please get in touch:

support@targetvalidation.org

Greyed our areas in the flower image above indicate there is no information for those data types, i.e. Affected Pathways, Animal models and Somatic mutations.

The genetic evidence that supports the association (*CD86* – multiple sclerosis) comes from the GWAS catalog and is processed by our pipeline.

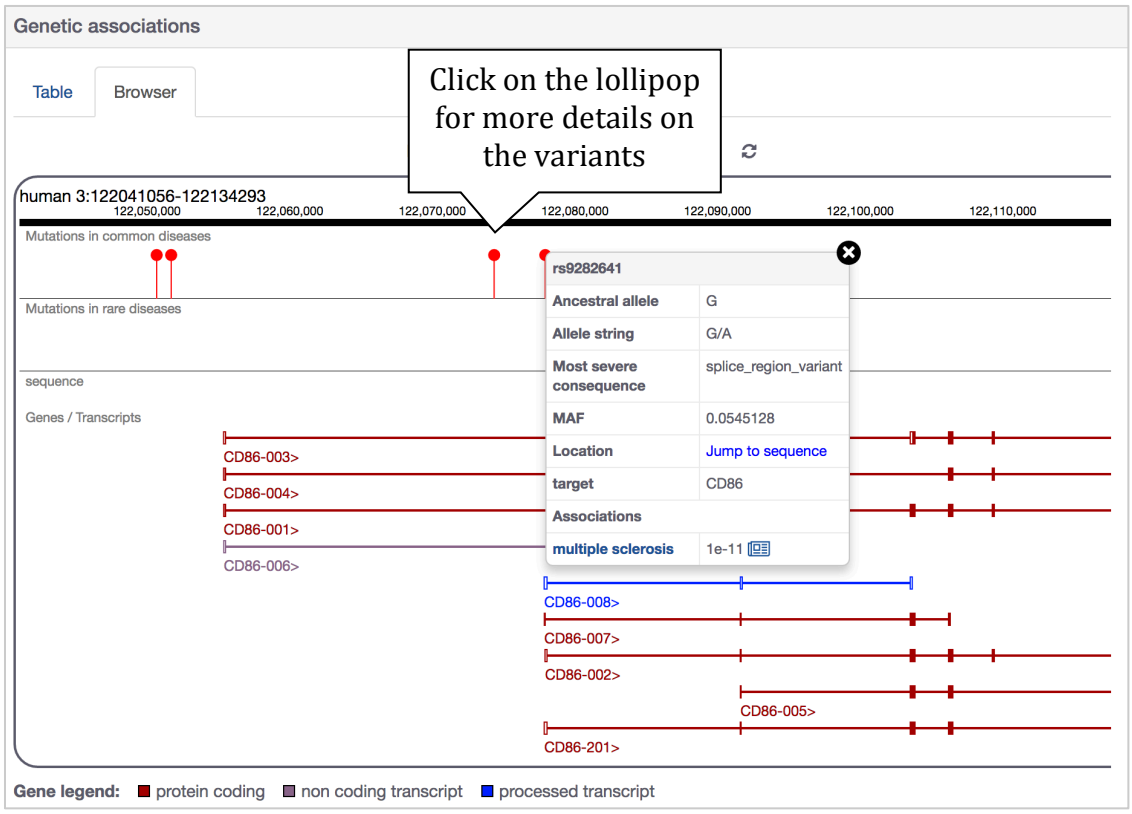
These variants are known in public databases i.e. dbSNP (the hint is on the rsIDs, such as rs9282641, rs4308217, rs2255214, and rs2681424.



In addition to the table above, you can also explore the 'Genetic associations' data in a Browser view.

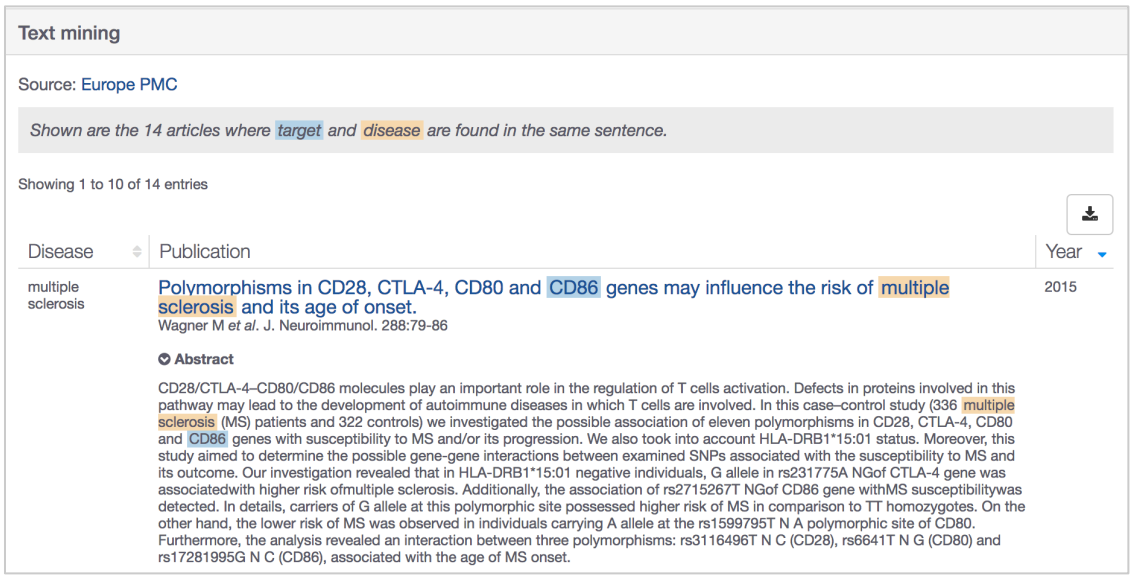
In the Browser view, you can zoom in and out, scroll along the genome and find out more about the gene (s), transcript (s), and the genetic variants (depicted as lollipops).

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

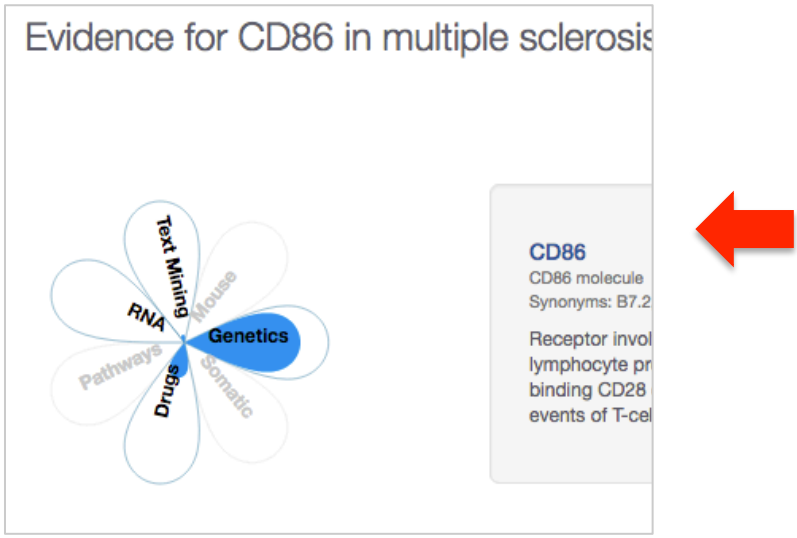


To find out if there are drugs currently in clinical trials targeting *CD86* in patients with multiple sclerosis, let's expand the 'Drugs' tab. There is just one drug (i.e. ABATACEPT) on phase II of clinical trials. There are two studies from clinicaltrials.gov, NCT01116427 (in multiple sclerosis) and NCT00035529 (in relapsing-remitting multiple sclerosis).

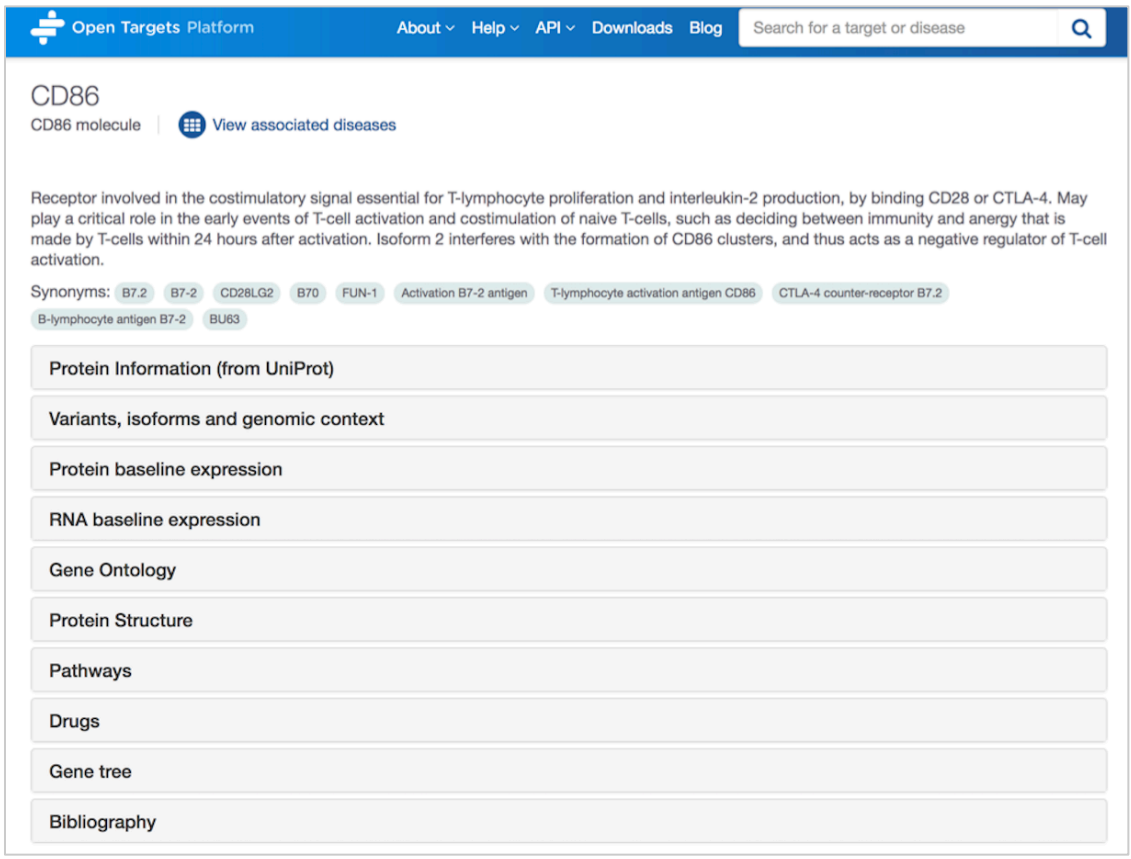
You may also want to find out the research articles that have been mined for the co-occurrence of gene name and disease in the same sentence:



We can now scroll back to the top of the page and click on the 'CD86 link (next to the flower) to explore this gene in more details, such as to find out its RNA expression levels across several tissues and whether or not it is associated with other diseases of the nervous system.



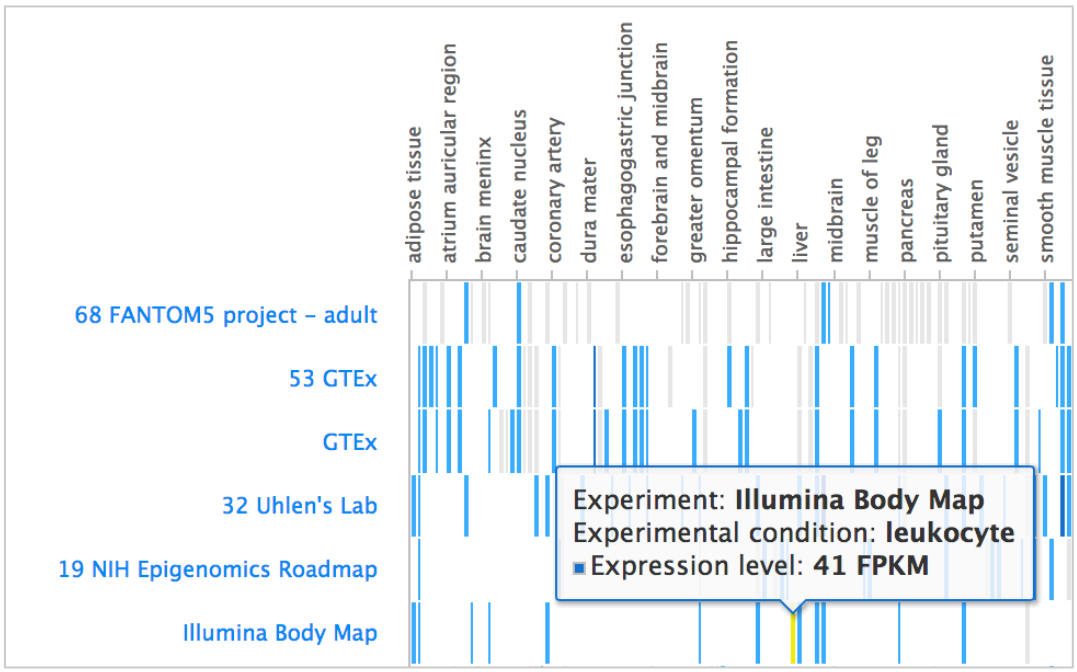
You will land on a page like this:



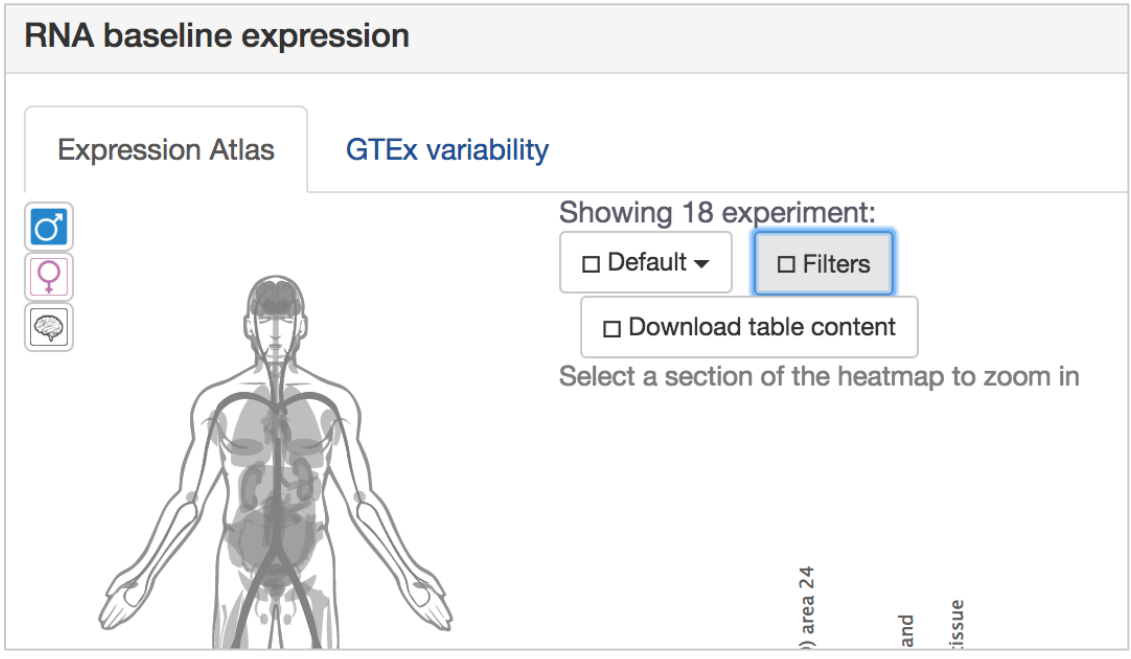
Expand the 'RNA baseline expression' to find out in which tissues *CD86* is highly expressed.

You will find two tabs in there: one with Expression Atlas data (with data from several projects including the Illumina Body Map), and one with GTEx data only.

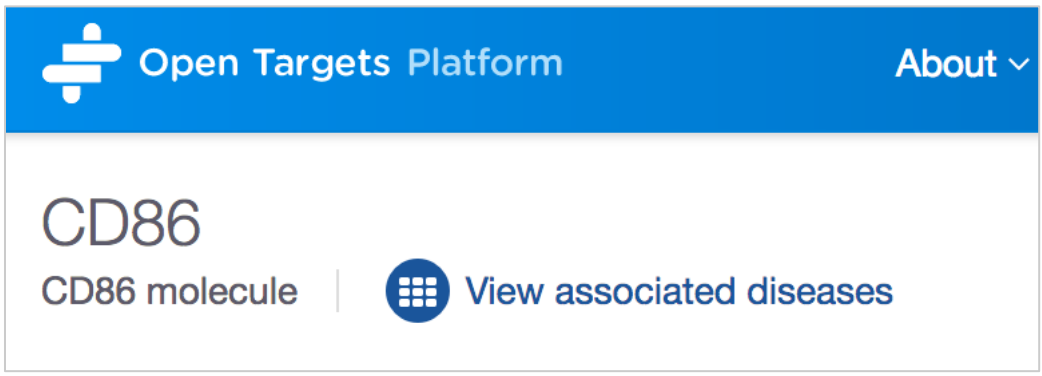
According to the Illumina Body Map data, leukocytes are the cells with the highest expression level for the *CD86* gene:



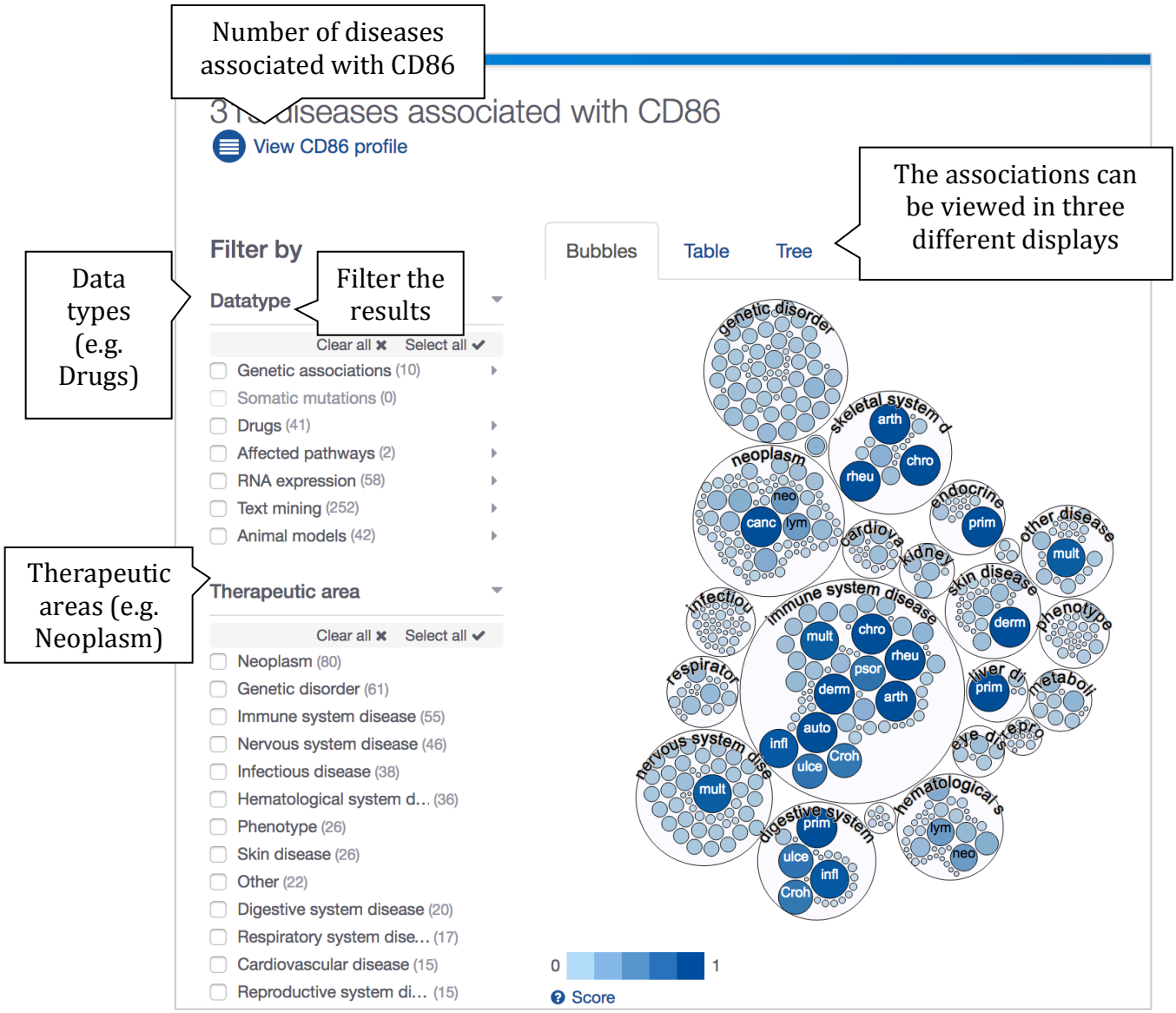
You can zoom this view in and out and apply filters such as 'Expression Value' (below cutoff, low, medium or high) and Anatomical Systems.



Let's now click on 'View associated diseases' to find out all diseases associated with the *CD86* (not only multiple sclerosis)*:*



You will land on a page like this:



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

- 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.

- Table view

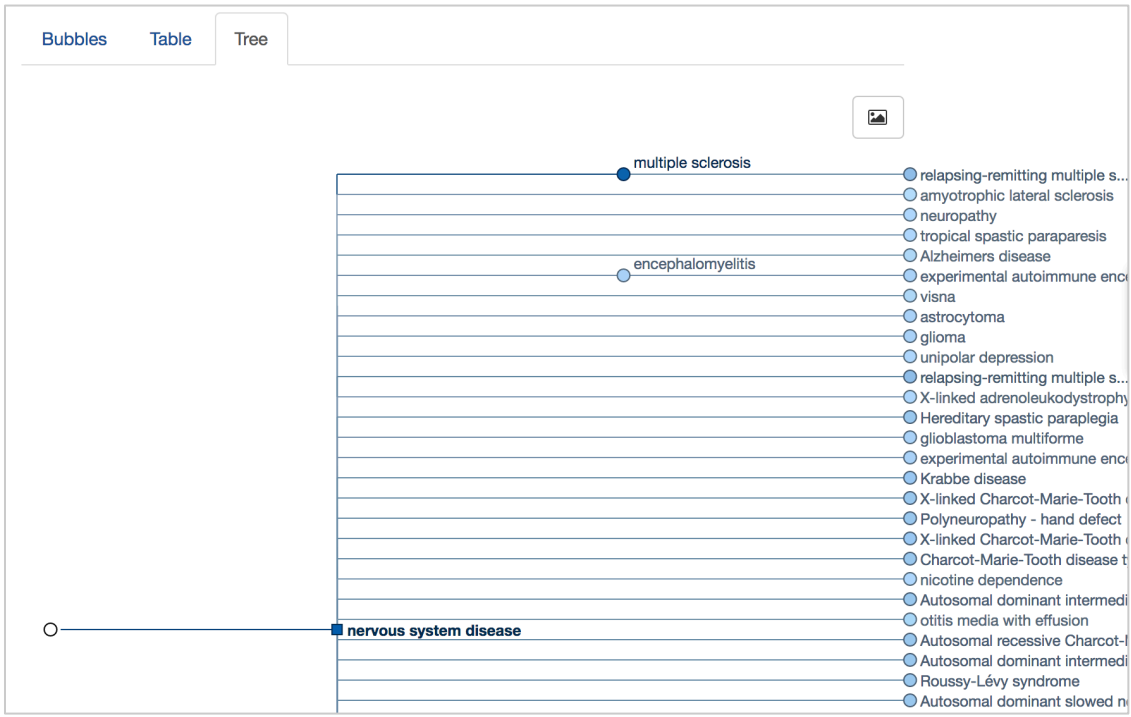
In this view, we list all diseases associated with a target, ordered by the association score, which is colour coded. When there is no evidence to support the association, the cells in this table are coloured in white (score of zero). You can show the 10 first entries and get the pagination for the remaining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

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

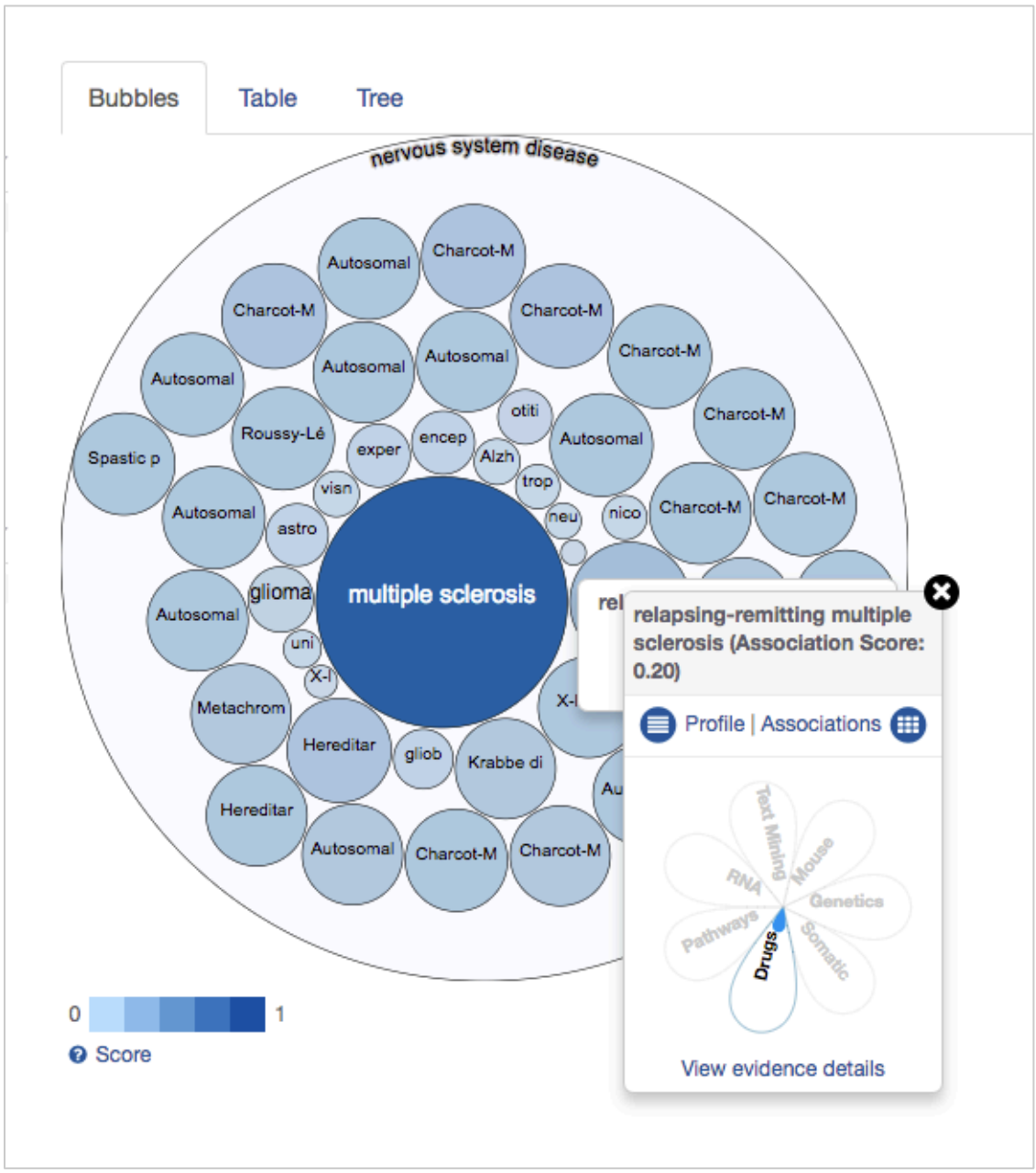
Tip: We colour code the cells in the table in different shade of blue as a visual way to convey the strength of the association (strongest association is coloured in dark blue). The score varies from 0 to 1. Hover over the cells in the table to view the numbers. Alternatively, you can select the cells in the table so that you can view the numerical values.

- Tree view

In the Tree view, you can visualise the evidence across the therapeutic areas in a tree format that represents the relationships of diseases. Therapeutic areas have a square symbol (e.g. Genetic disorders), while the diseases (e.g. ovarian carcinoma) are represented as circles. The squares and circles are colour coded in blue, and the darker the blue, the stronger the association:



Regardless the view you choose to explore, you can filter the data to find out if there are other “Nervous system” diseases associated with this target. There are 46 of them (e.g. Hereditary spastic paraplegia, amyotrophic lateral sclerosis and neuropathy). The scores vary from 0.01 to 0.20 (score of 0.89 for multiple sclerosis).


Bubbles
Table
Tree
nervous system disease
multiple sclerosis
Autosomal
Charcot-M
Roussy-Lé
Spastic p
exper
encep
Alzh
otiti
trop
neu
nico
visn
astro
glioma
uni
X-l
Metachrom
Hereditar
gliob
Krabbe di
0
1
Score
relapsing-remitting multiple sclerosis (Association Score: 0.20)
Profile | Associations
Text Mining
Mouse
RNA
Genetics
Pathways
Drugs
Somatic
View evidence details

HANDS-ON EXERCISES PART I

Exercise 1

TRAF3IP2 in psoriasis

BACKGROUND
The *TRAF3IP2* gene encodes a protein that is involved in regulating responses to cytokines by members of the Rel/NF-kappaB transcription factor family. These factors play a central role in innate immunity in response to pathogens, inflammatory signals and stress. Genome-wide association studies have identified a psoriasis susceptibility locus at *TRAF3IP2* (PMID:20953188, PMID: 20953186).

QUESTIONS
a) Is this target associated with other skin diseases? If so, can you name three of them with the highest (overall) association score (> 0.70)?

b) Which types of data suggest such associations?

c) Can you name a few genetic variants that support the association between *TRAF3IP2* and psoriasis? Where does this evidence come from?

d) Can you view the genetic variants in a Browser view? What does the different colours represent in this view? In which human chromosome does this gene map to on the human genome? Is it a forward or reverse stranded gene?

e) How many research articles do currently support the association between *TRAF3IP2 in psoriasis*?

Exercise 2

Aducanumab, antibody that may slow the progress of Alzheimer's disease

BACKGROUND

The most common cause of dementia, Alzheimer's disease is a brain disorder that leads to impairment of intellectual functions and social skills, resulting in significant disturbance of daily living activities. Dementia is caused by the destruction of cells in certain parts of the brain primarily involved in memory and mental function.

SIGNIFICANCE

Alzheimer's disease is the 6th leading cause of death in the United States and it's expected 16 million Americans will be living with Alzheimer's by 2050 (source: alz.org).

Biogen in collaboration with Neurimmune has been conducting clinical studies investigating a potential new treatment (Aducanumab) for early Alzheimer's disease. Aducanumab is an antibody that binds to and may reduce amyloid plaques from the brain, potentially slowing the progress of the disease.

QUESTIONS

a) Which gene (s) does this drug match to?

b) Can you name the biochemical pathways that this gene is involved in and that likely play a role in the pathogenesis of Alzheimer's?

c) Are there other pathways that this gene can be mapped to, outside the context of any disease? Hint: these pathways are not related to Alzheimer's or other diseases. Click on the gene name to get to the profile page of the target (or gene).

d) What is the human tissue with the highest RNA baseline expression for this target according to GTEx? Can you compare the RNA expression with the protein expression in this same human tissue?

e) Is there a homologue to the human gene targeted by aducanumab in zebrafish? *Note: Zebrafish can be a useful animal model to focus on the roles of intracellular aggregate-prone proteins in the pathogenesis of AD and to help the identification of pathways, which can enhance the clearance of these toxic proteins. See more in the link below:*

http://www.pdn.cam.ac.uk/directory/angeleen-fleming

Exercise 3

BACKGROUND
Franke et al (2006) described seven genes associated with Alzheimer's disease:

HFE
PSEN1
TF
APOE
ADRB2
PSEN2
A2M

The information in Open Targets will help you to choose and prioritise the best targets from that list.

Can you input this data as a list (in .txt) and filter the associations for Alzheimer's focusing on those seven genes only? Yes, you can with the 'Your target list' filter in the association page for Alzheimer's. Scroll down in that page to find 'Your target list' and upload your text file.

QUESTIONS
a) Which of those seven genes have the strongest association w/ Alzheimer's?

b) Are there any targets, which are membrane receptors?
We can now use the filter 'Target class' to focus on membrane receptors only. There is only one target that is a membrane receptor, *ADRB2*.

c) Which amino acids of this membrane receptor correspond to the extracellular domain?

HANDS-ON EXERCISES PART II

Exercise 4

IL6 as a possible target in the treatment of rheumatoid arthritis

BACKGROUND
IL6 encodes a cytokine that functions in inflammation and the maturation of B cells. In addition, it has been shown to be an endogenous pyrogen capable of inducing fever in people with autoimmune diseases or infections. The protein implicated in a wide variety of inflammation-associated disease states, including systemic juvenile rheumatoid arthritis (PMID:9769329).

QUESTIONS
a) How many diseases within the broader Therapeutic area 'Immune system disease' are associated with this target based on 'Drugs' as the only class of data type?

b) In addition to the data type 'Drugs', are there other types of evidence supporting the association between *IL6* and rheumatoid arthritis?

c) Can you describe some of the phenotypes in mice that seem to mimic the phenotype in patients with rheumatoid arthritis? What are the differences between the existing *Il6* knockout mouse models for this disease?

Let's now explore some disease information available for rheumatoid arthritis.

d) What are the different subtypes (i.e. the children terms) of rheumatoid arthritis (RA)? Can you download the ontology tree of rheumatoid arthritis? Let's now have a look at one of the subtypes of RA i.e. chronic childhood arthritis.

e) Can you list all drugs in phase III of clinical trials for the treatment of chronic childhood arthritis, which are still recruiting volunteers?

Let's now click on 'View associated targets' to see the targets associated with chronic childhood arthritis.

f) Can you filter this list to show the results for the following targets only: *IL1B*, *JAK2*, *CXCR4*, *TNF* and *PTGS2* only? Are any of these genes a secreted class of target?

Exercise 5

BACKGROUND
The Drug Discovery Unit at the University of Dundee translates biology research into novel drug targets and candidate drugs to address unmet medical need across, such as infectious affecting mainly the developing world.

These are some of the targets that have been selected for further investigation: *ODC1*, *GUCY1A2*, *GUCY1A1*, *CSF3R*, *CCR5*, *TLR7*, and *CD28*.

QUESTIONS
a) How can you search for all these targets in one go using the web interface of the Open Targets Platform?

b) Which diseases do these targets map to?

c) Can you list a few of the pathways these targets are involved in? Can you find out the diagram of the pathway that is enriched in this list of targets?

d) Can you list a few of the drugs targeting some of these genes? Can you find out the mechanism of action for one ACETAMINOPHEN?

e) Do any of these targets seem to be interacting with each other?

QUICK GUIDE TO DATABASES

Here is a list of databases and projects that may be useful for you to explore:

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 Trust 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/

IntOGen - It provides evidence of somatic mutations, genes and pathways involved in tumorigenesis 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

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/

ANIMAL MODELS

Phenodigm - The Phenodigm resource at the Wellcome Trust Sanger Institute provides evidence on associations of targets and disease. It uses a semantic approach to map between clinical features observed in humans and mouse phenotype annotations.

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

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/