Open Targets: mining gene-disease associations for drug discovery

Alzheimer's Research UK – Oxford DDI

23rd October 2017

Denise Carvalho-Silva
Scientific Outreach Lead
Open Targets

Notes

This workshop is based on the June release of the Open Targets Platform.

Some useful links:

1) Science in Open Targets
https://www.opentargets.org/science/

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

3) Platform FAQs
https://www.targetvalidation.org/faq

Questions or Feedback?

support@targetvalidation.org

TABLE OF CONTENTS

OVERVIEW

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

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 Biogen (https://www.biogen.com/) joined the initiative. The consortium was rebranded to Open Targets in April 2016.

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
- Introduce 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 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 biological 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 biological data to facilitate the identification and selection of targets for new therapies.

We use biological 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, and Animal models.

This evidence is used for the ranking and scoring of target-disease associations, which depends on the frequency of evidence, the confidence and severity. We then aggregate the evidence 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 latest release of the Platform (June 2017) contains:

- 26,122 targets
- 9,150 diseases
- 2,857,732 associations
- 5,347,817 evidence

In addition to the web interface of the Platform, our data can be accessed programmatically with the REST API. Dump files with all associations and evidence are also available for download.

The Open Targets Platform is for all types of users (biologists and otherwise), from both academia and industry. Our users can browse a target on a gene by gene (or disease by disease) basis, search for a batch of up to 200 genes (Batch search), 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 which targets are associated with a disease,
- Find which diseases are associated with a target,
- Find the evidence supporting target-disease associations,
- Find the associations of a target with diseases from different therapeutic areas,
- Find target specific information in our target profile page,
- Find disease specific information in our disease profile page,
- Export data for target-disease association,
- Search for many targets at once with our batch search tool.

Connect with us

❖ Open Targets Blog

❖ Follow us on Twitter

- Check our page on Facebook and LinkedIn
- Email us

How to cite us

Koscielny, G. *et al.* Nucleic Acids Res (2017 Database Issue) Volume 45, D985–D994

OPEN TARGETS PLATFORM: LIVE DEMOS

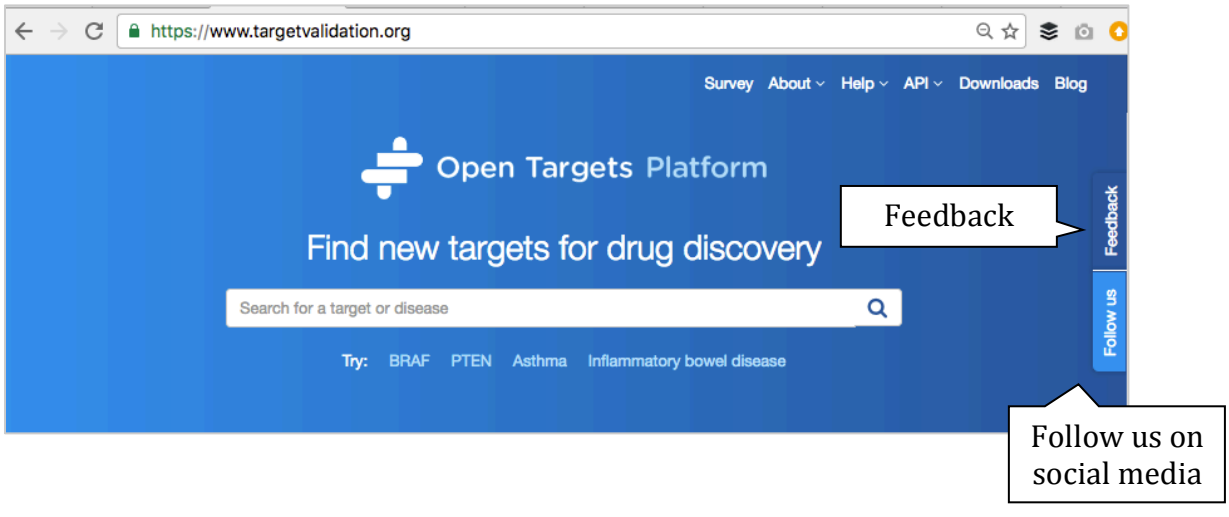
We will guide you through the website using multiple sclerosis, as an example and exploring the *CD86* gene.

The following points will be addressed during the walkthrough:

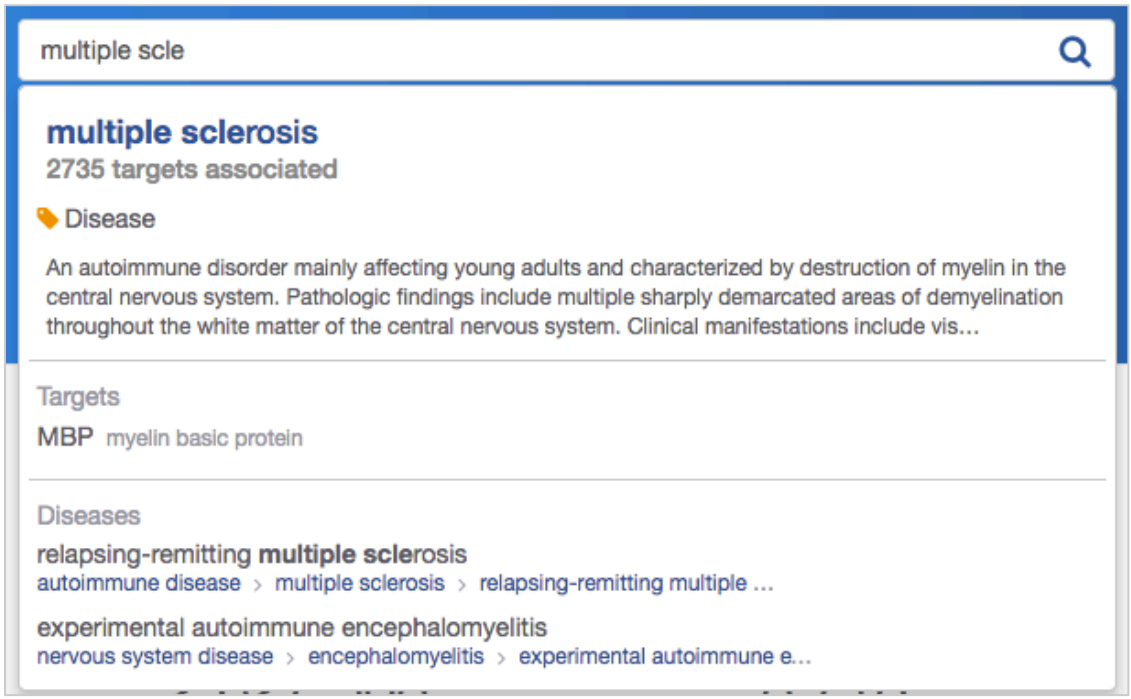
- How to find targets associated with multiple sclerosis
- How to filter down the number of targets based on specific type of evidence
- How to find the data sources used to support the target-disease association
- How to look for other diseases (than multiple sclerosis) associated with a given target
- 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

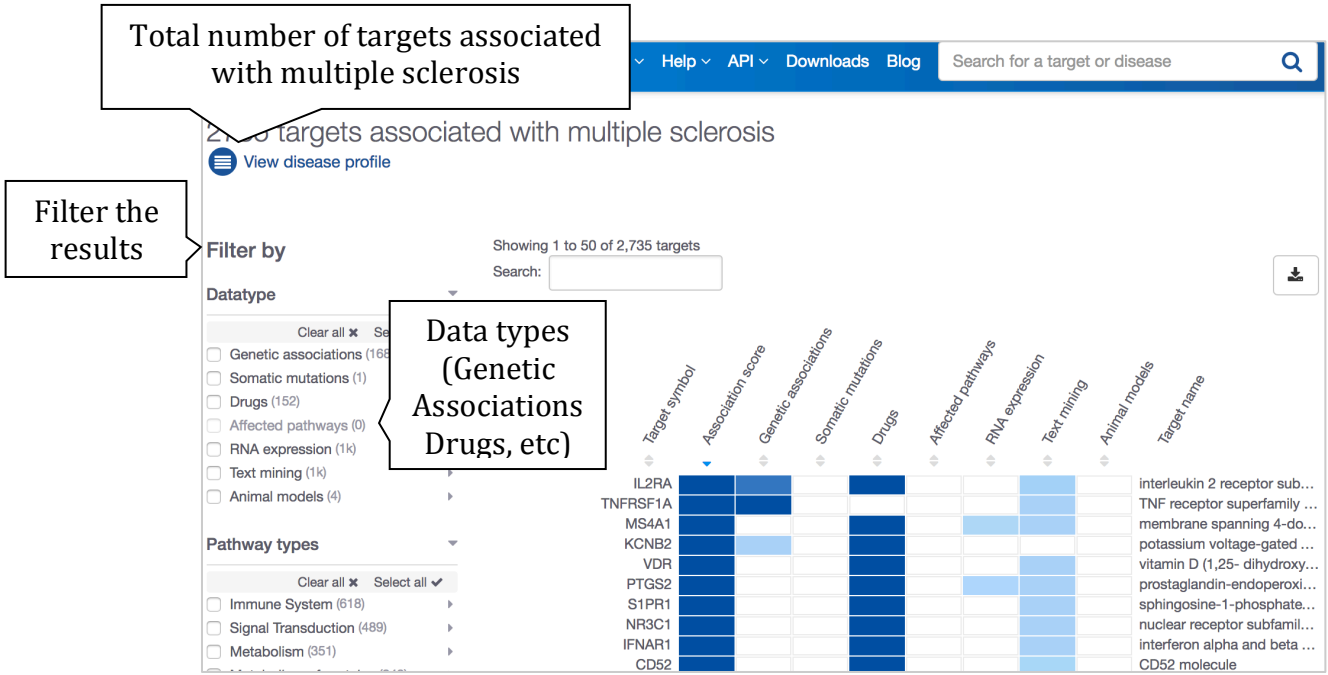
Go to www.targetvalidation.org and search for multiple sclerosis.



Select the first (best) hit:



You will see a page like this:



The current release of the Open Targets Platform (June 2017) lists 2906 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**?

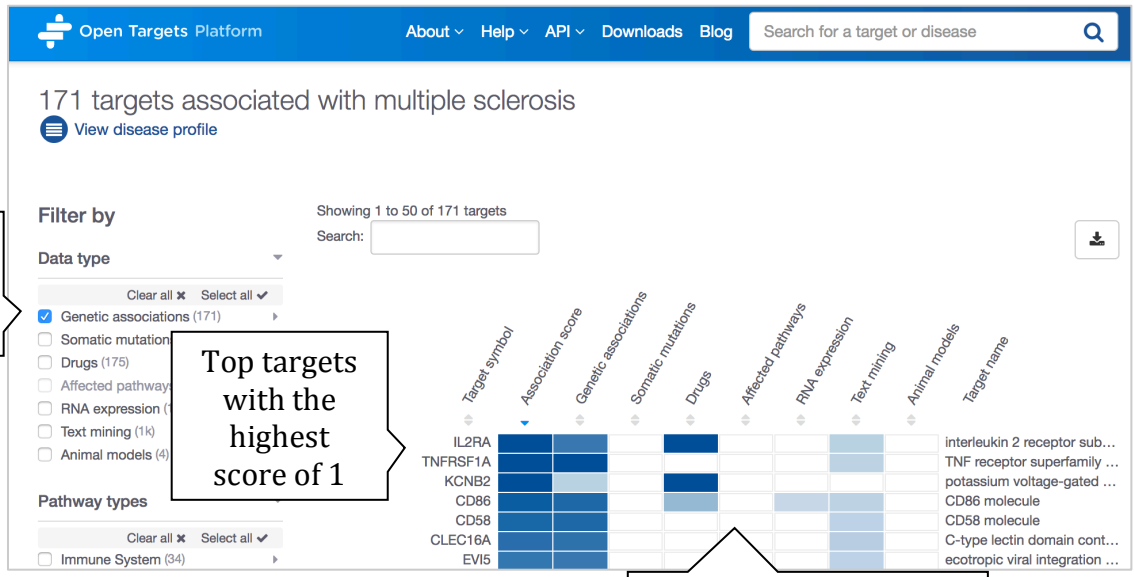
We collect data from various sources and combine them into categories called Data types. Example of data sources are GWAS catalog and UniProt, both combined into Data types. Note that data from an individual source can contribute to different Data types, e.g.

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:



Genetic Associations selected

Top targets with the highest score of 1

Blank cells have no data to support the - association (score of 0)

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.

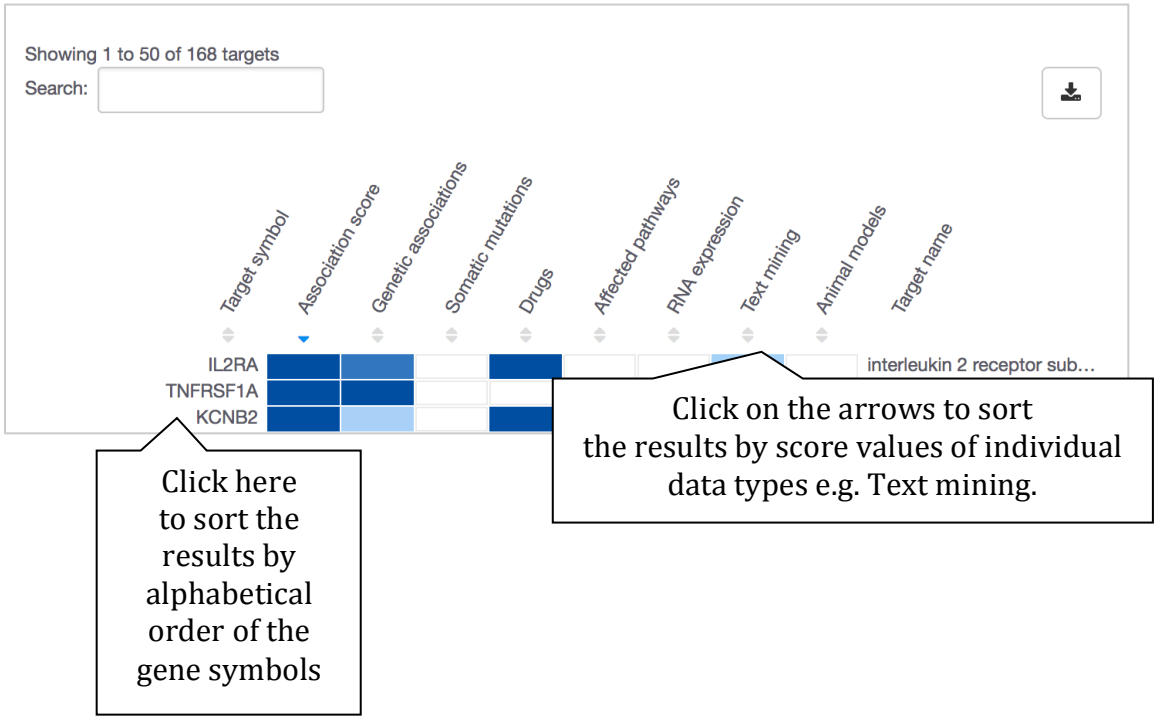
The table is sorted by default with the best hit on the top of the table i.e. *IL2RA*. The best hit is the target that contains the highest (overall) association score. Different weight is given to different data types when computing the score:

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

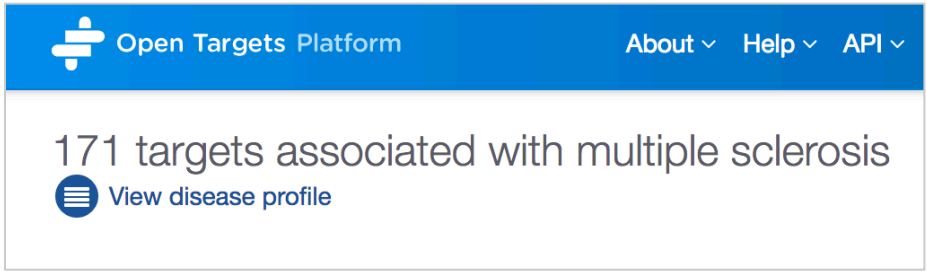
Genetic association = somatic mutations = drugs = pathways

RNA expression
Animal models = Text mining

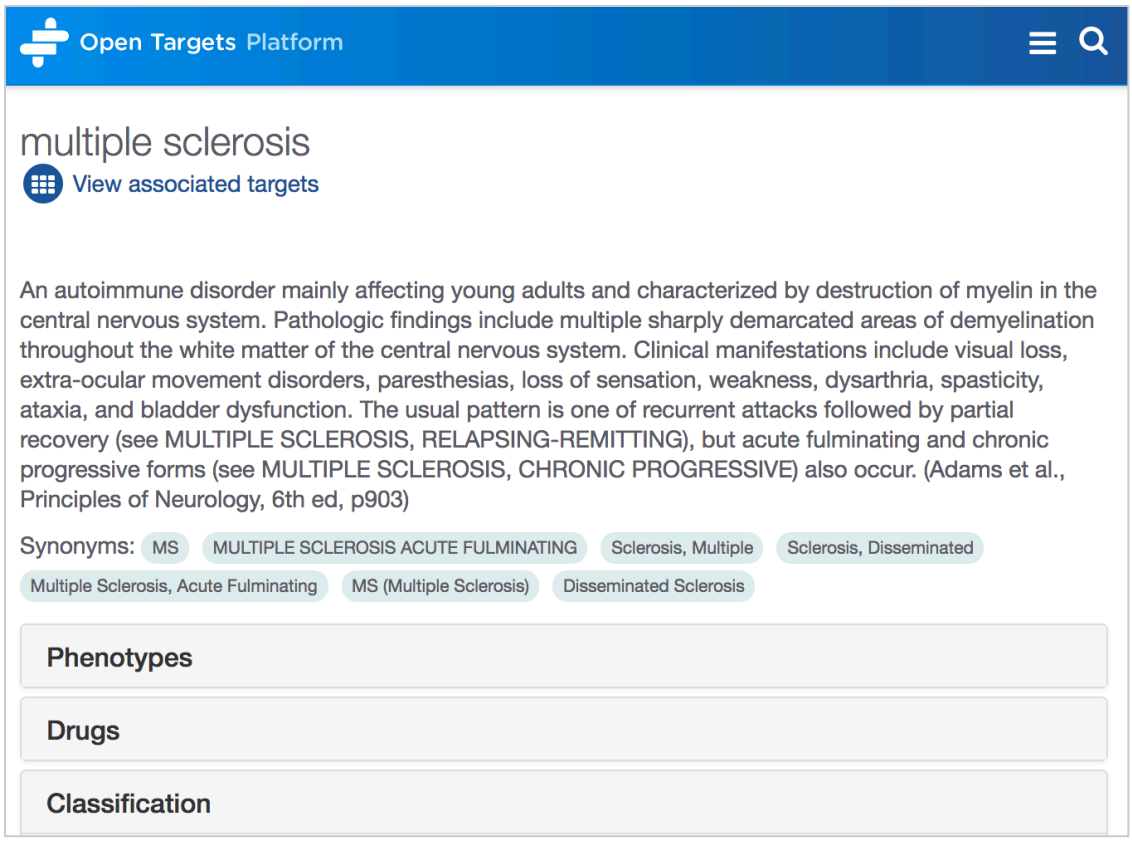
You can sort the table by alphabetical order of the list of targets, or by the association score (either overall or per data type e.g. Genetic associations, Drugs, Text mining, etc). The association score varies from 0 to 1, the closer to 1 the stronger the association. This score is computed for each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

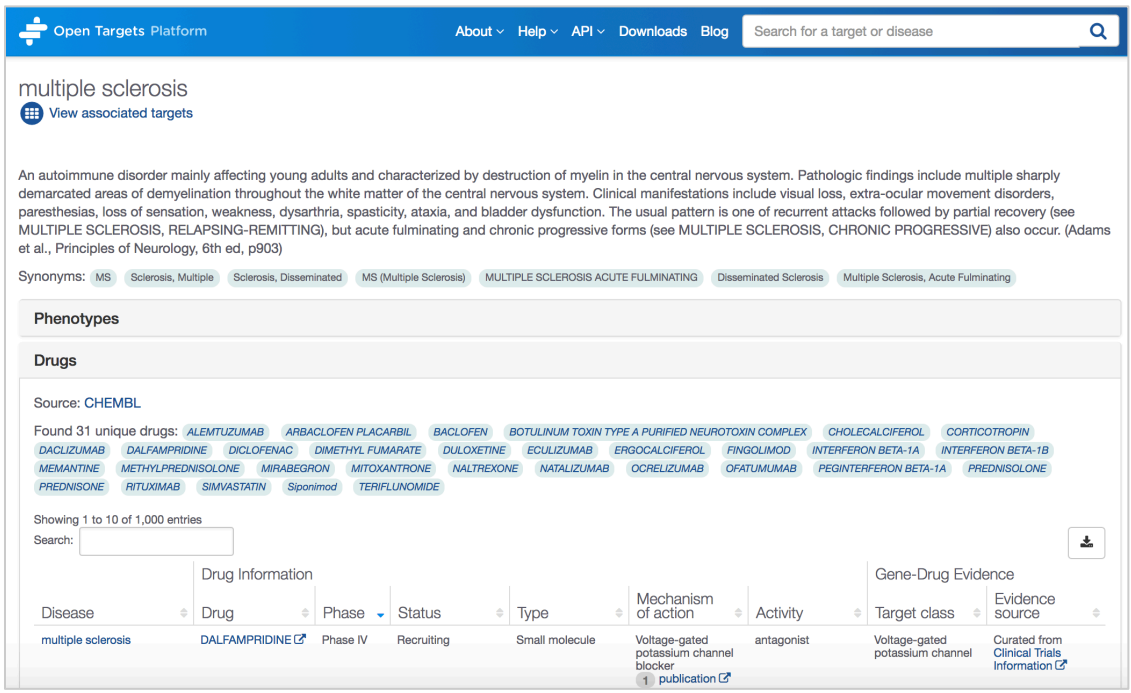


Let's now expand the tab 'Drugs'. You will see a page like this:

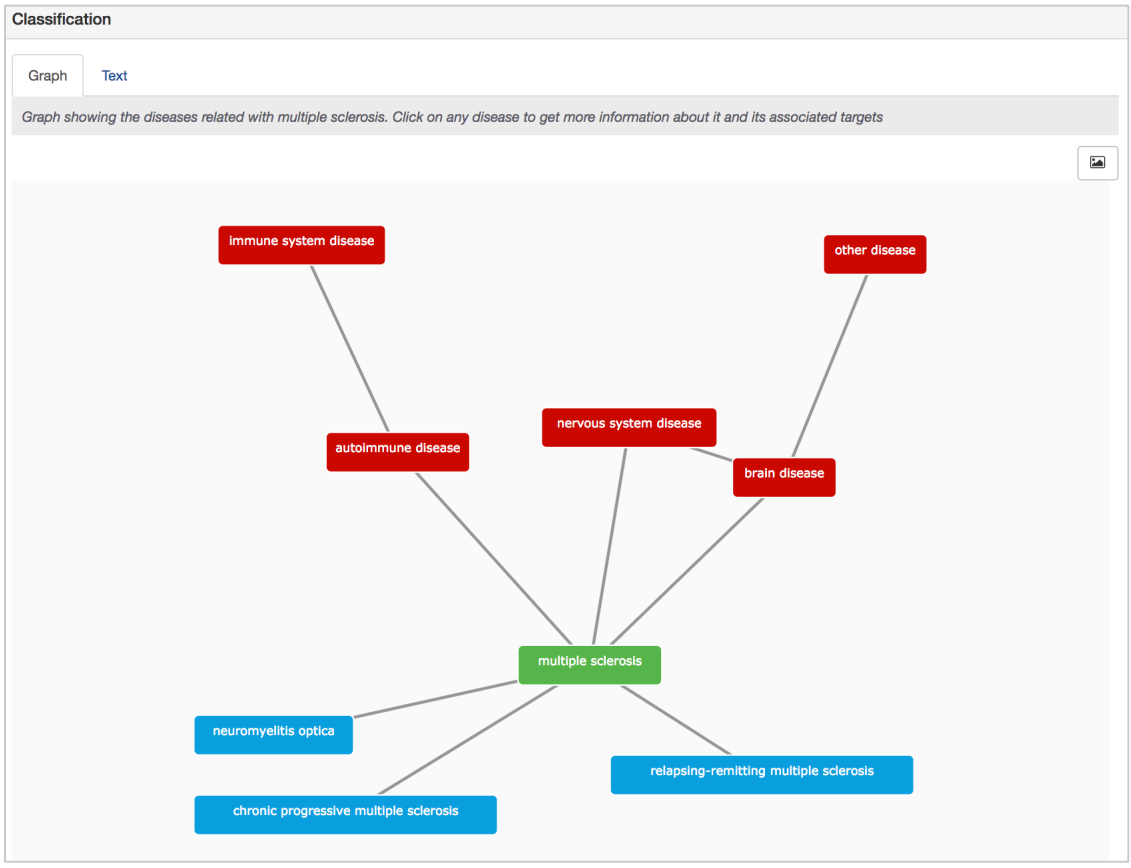


In the June 2017 release of the Open Targets Platform, we have 31 unique drugs in different phases of clinical trials with patients suffering from this condition. They will be targeting different proteins.

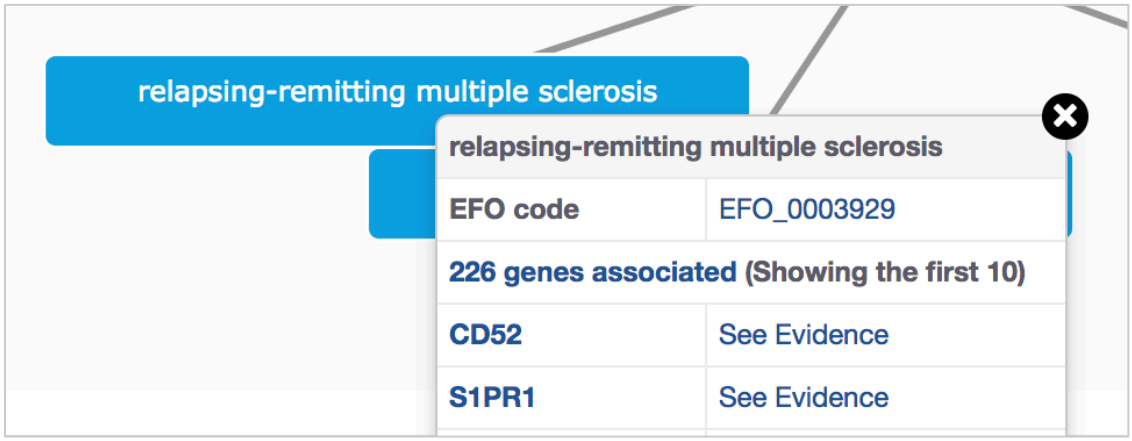
You can filter (and sort) the table by disease, phase of clinical trial (e.g. IV, the most advanced phase), class of the target (e.g. membrane receptor), and much more. You can also download this table in csv (comma separated value):



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'.

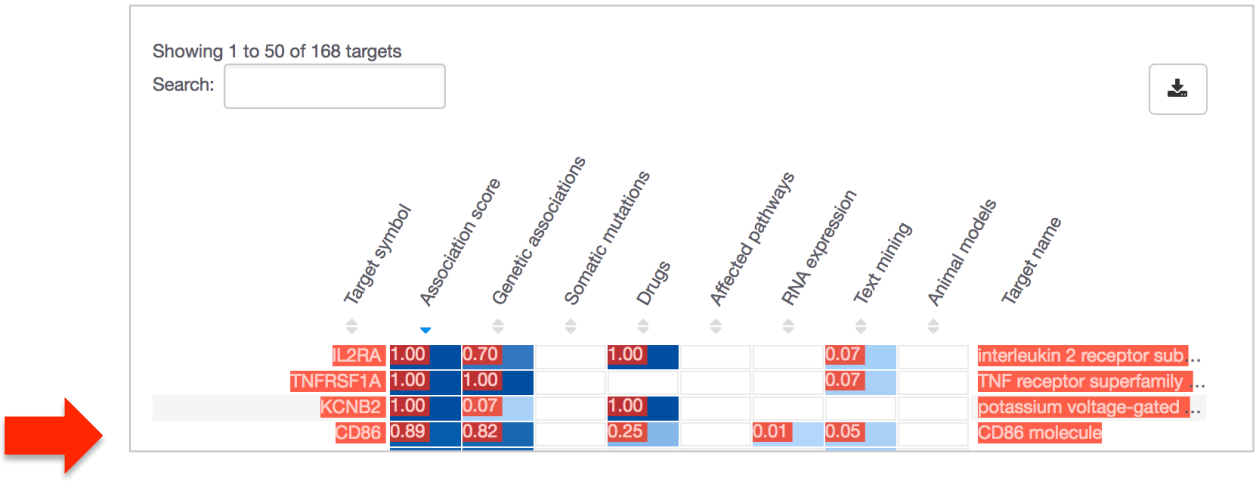
So you will get direct and indirect associations of diseases and genes. For more on this, check our blog post:

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.

If we click on any row corresponding to CD86 in the table below:



we will 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

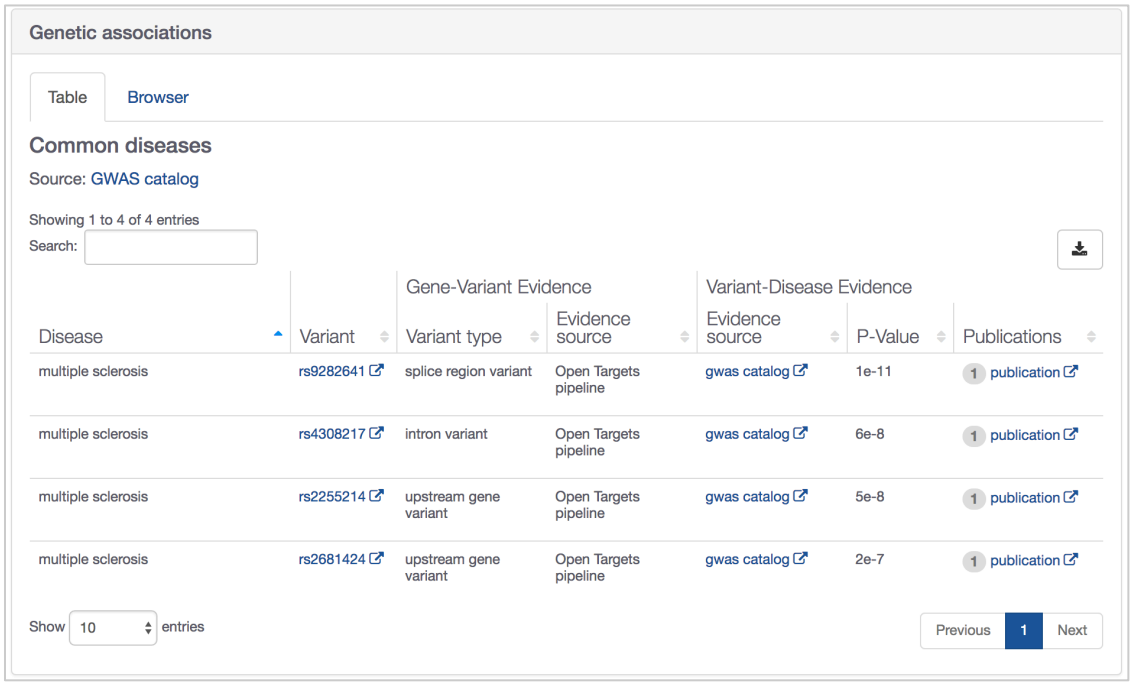
Note: If you wish to suggest data types or resources we could/should incorporate into our Platform, please get in touch:

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.

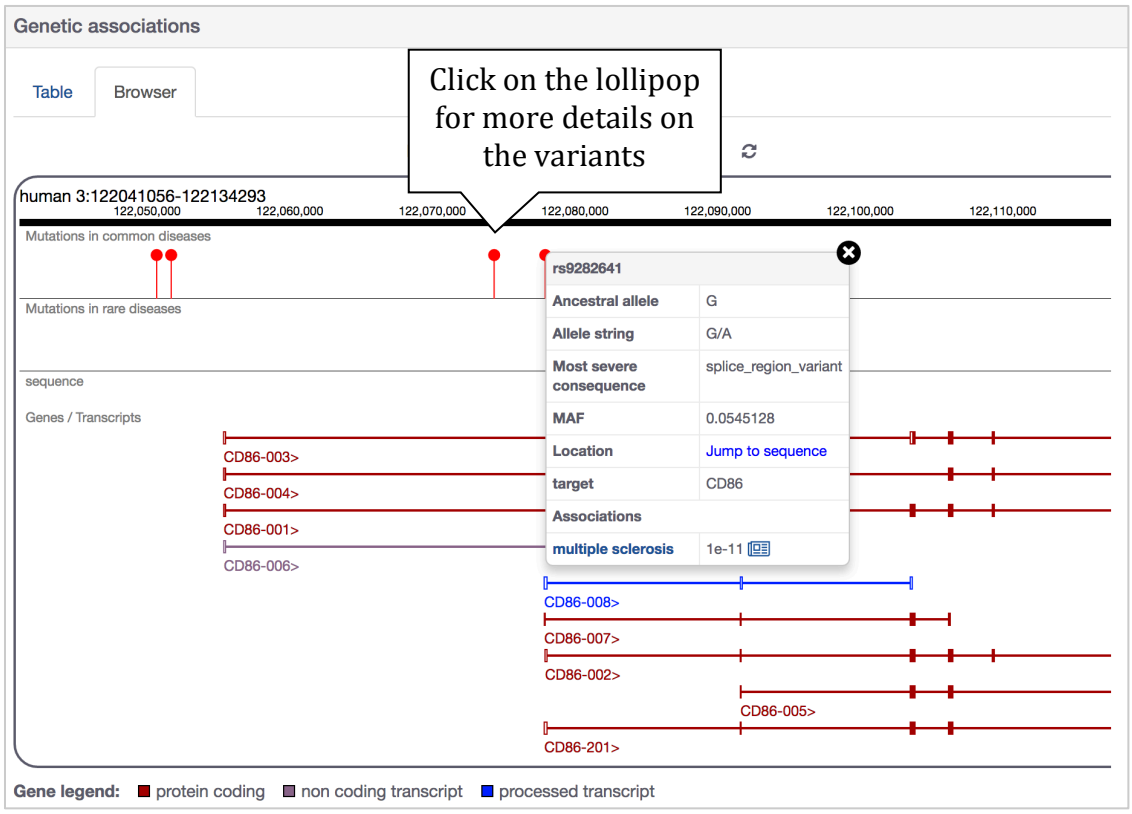
These variants are known in public databases i.e. dbSNP (the hint is on the rsIDs, such as rs9282641, rs4308217, rs2255214, and rs2681424.



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

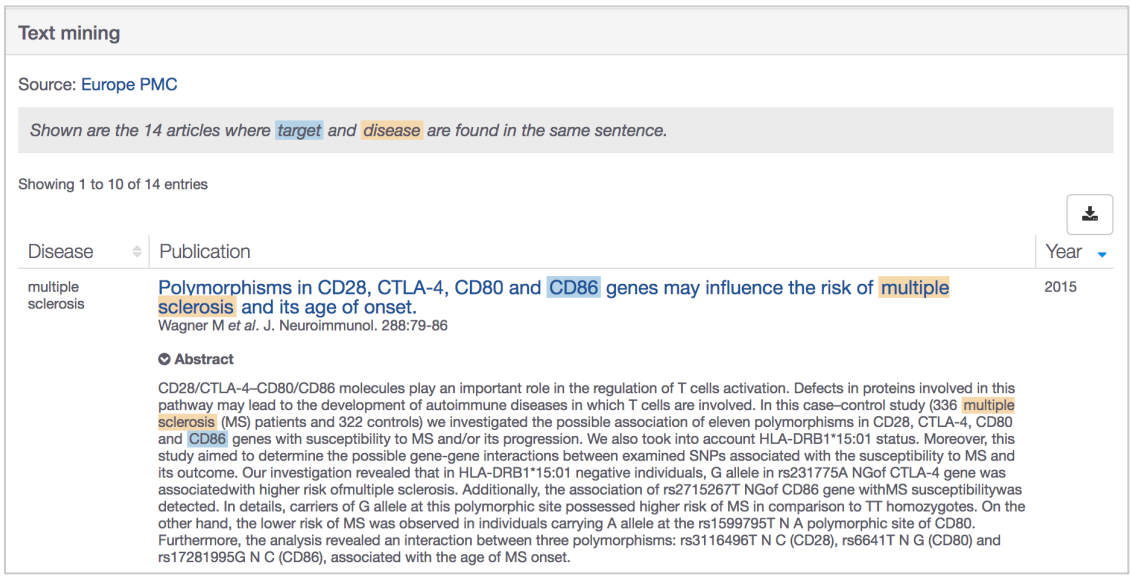
In the Browser view, you can zoom in and out, scroll along the genome and find out more about the gene (s), transcript (s), and the genetic variants (depicted as lollipops).

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

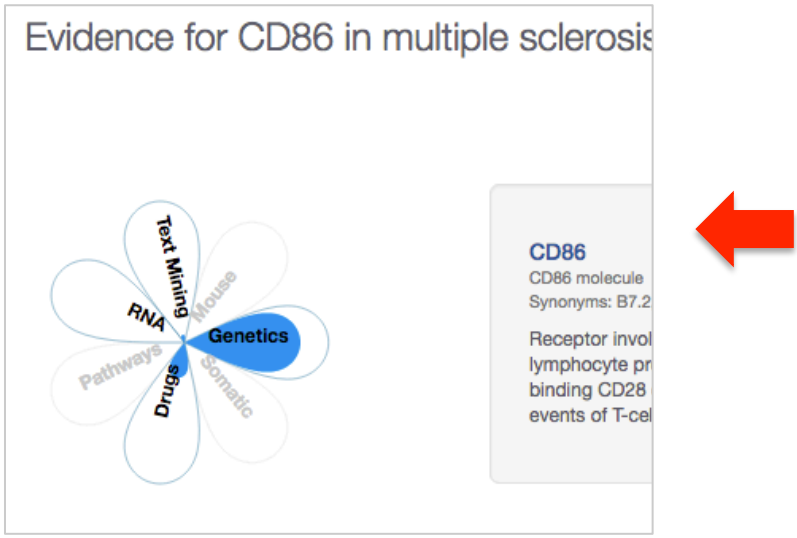


To find out if there are drugs currently in clinical trials targeting *CD86* in patients with multiple sclerosis, let's expand the 'Drugs' tab. There is just one drug (i.e. ABATACEPT) on phase II of clinical trials. There are two studies from clinicaltrials.gov, NCT01116427 (in multiple sclerosis) and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

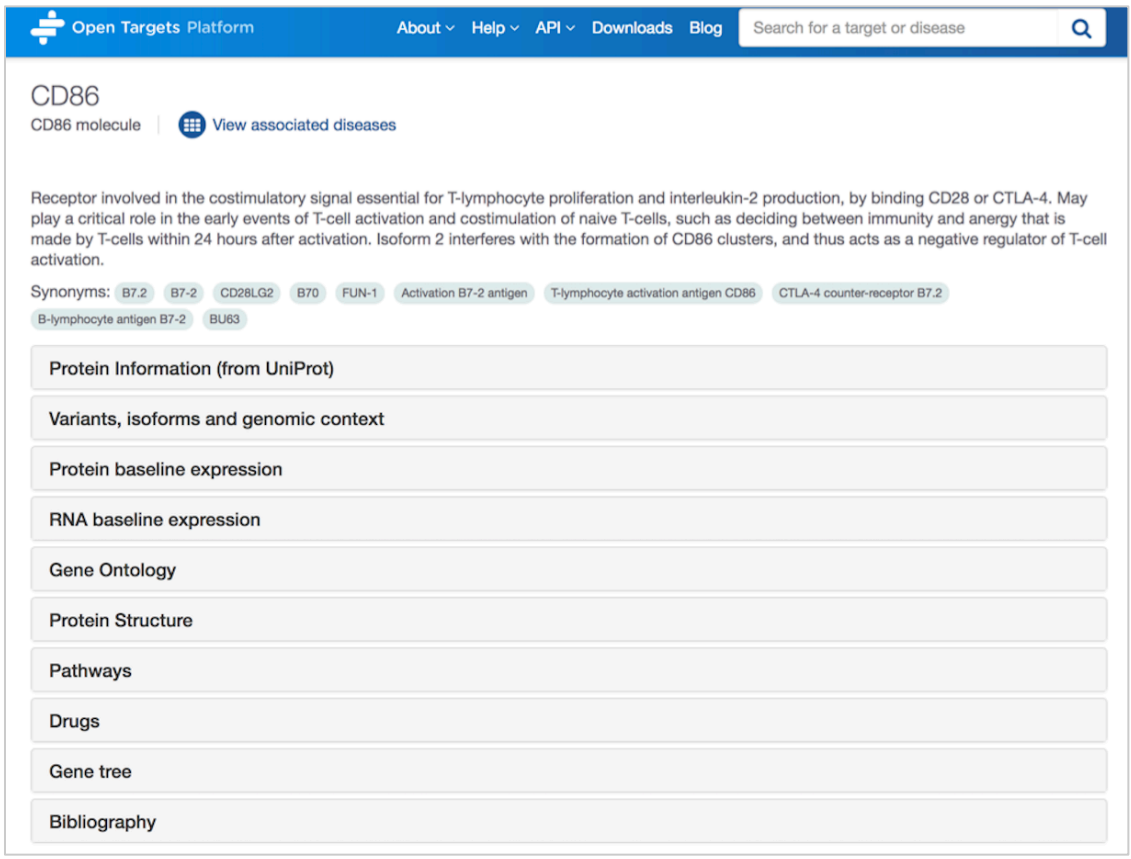
You may also want to find out the research articles that have been mined for the co-occurrence of gene name and disease in the same sentence:



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.



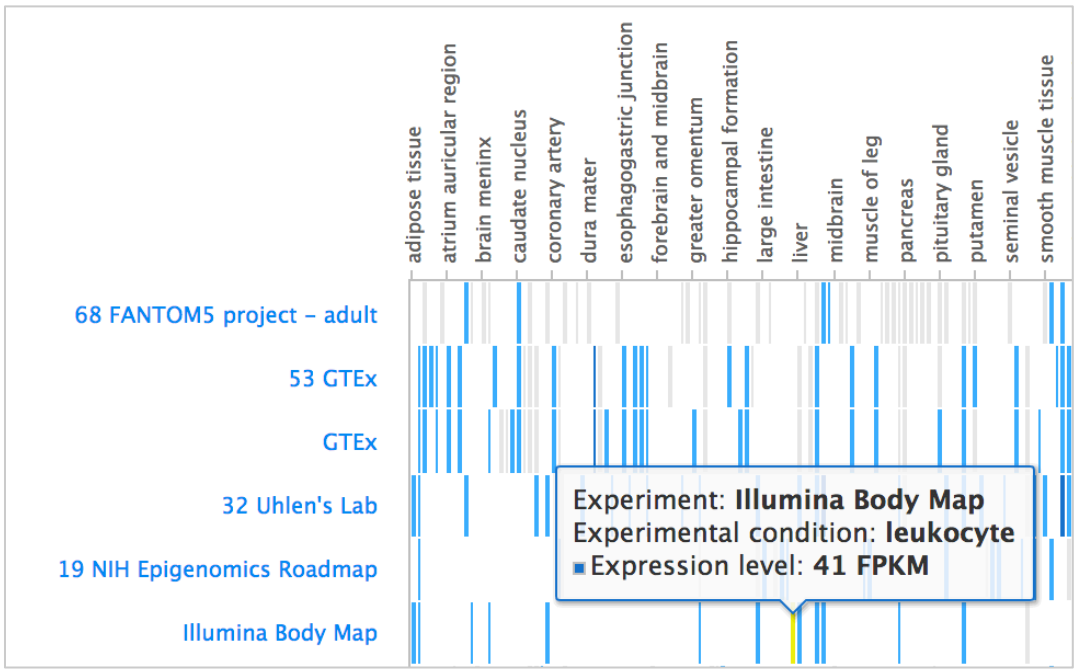
You will land on a page like this:



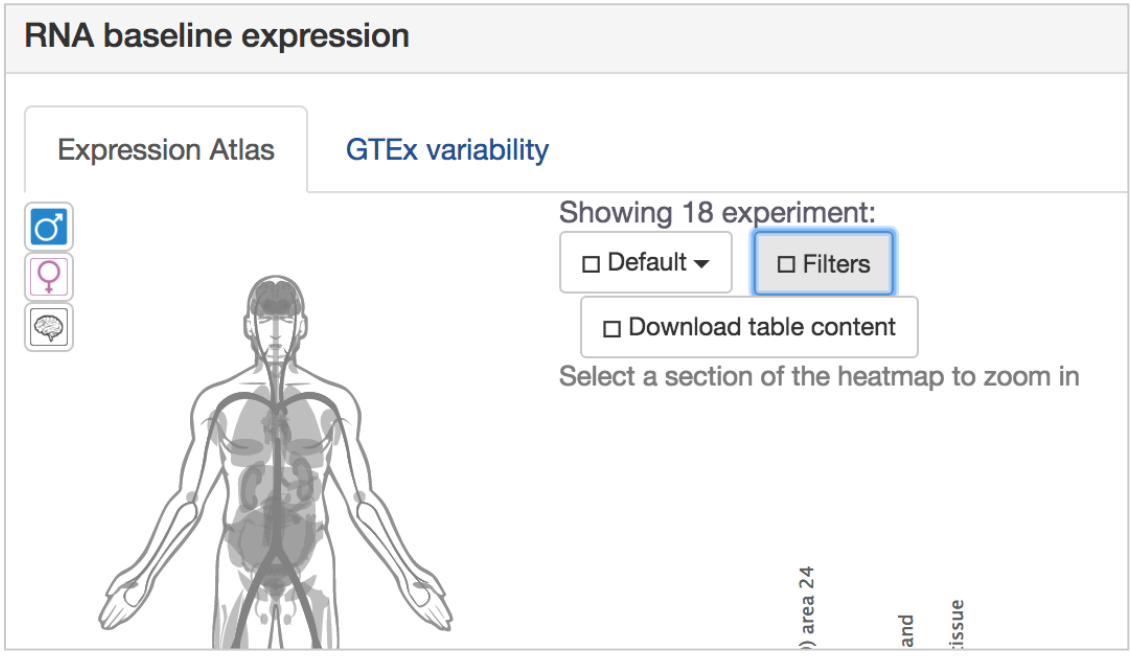
Expand the 'RNA baseline expression' to find out in which tissues *CD86* is highly expressed.

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.

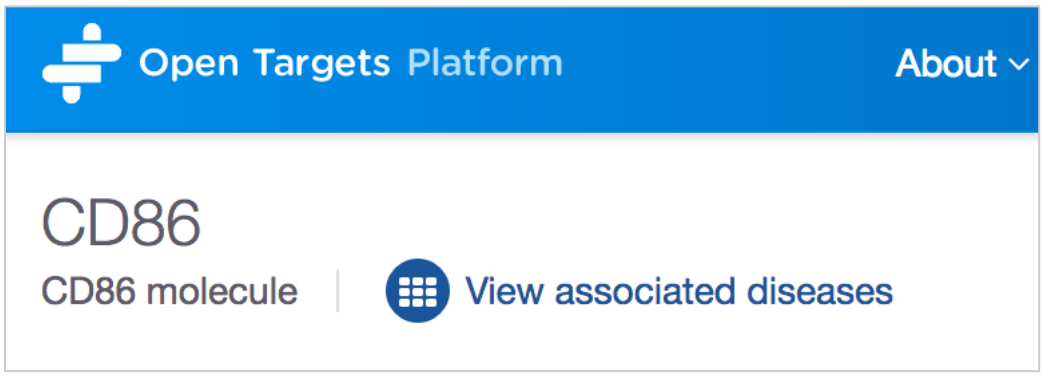
According to the Illumina Body Map data, leukocytes are the cells with the highest expression level for the *CD86* gene:



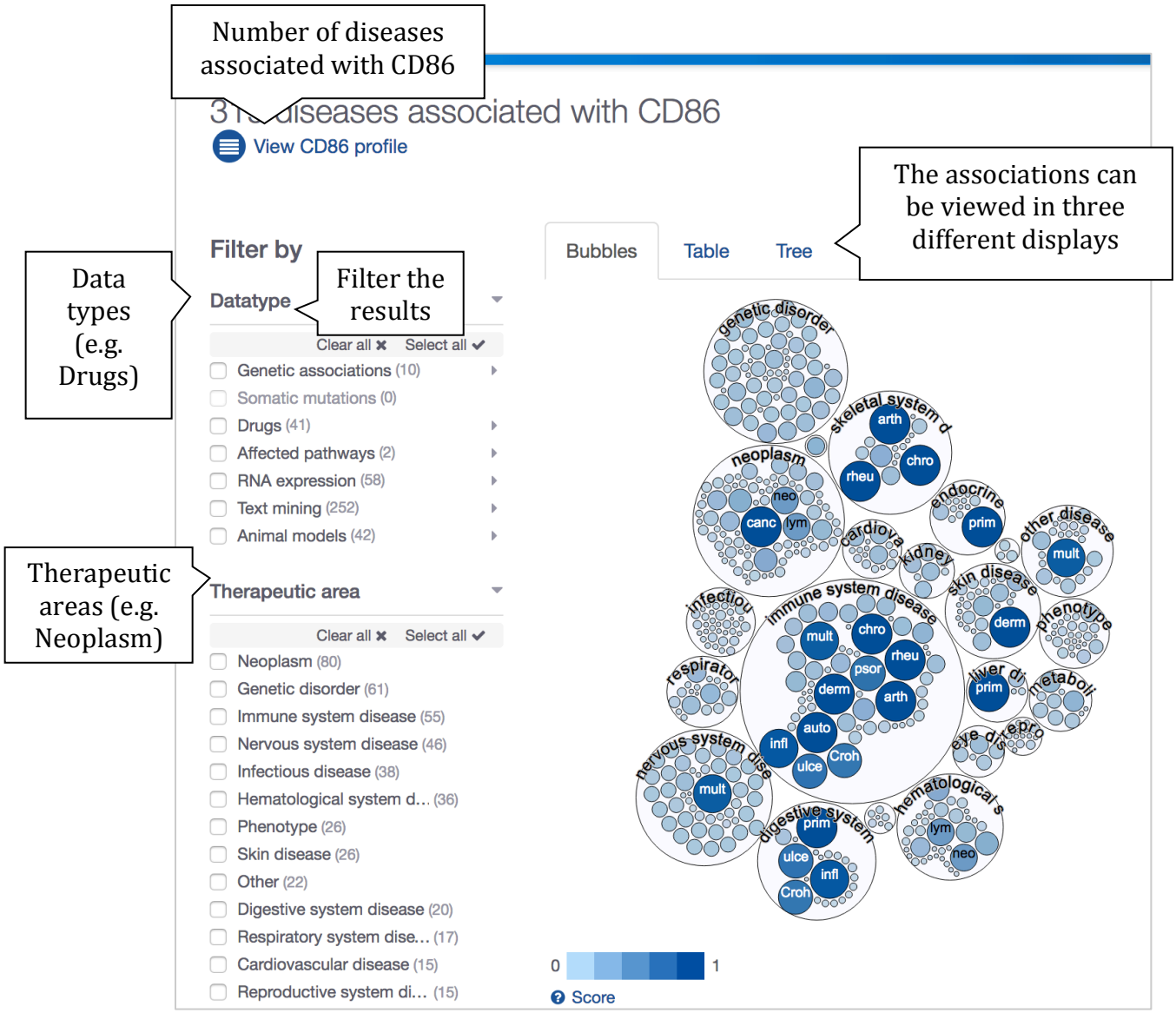
You can zoom this view in and out and apply filters such as 'Expression Value' (below cutoff, low, medium or high) and Anatomical Systems.



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

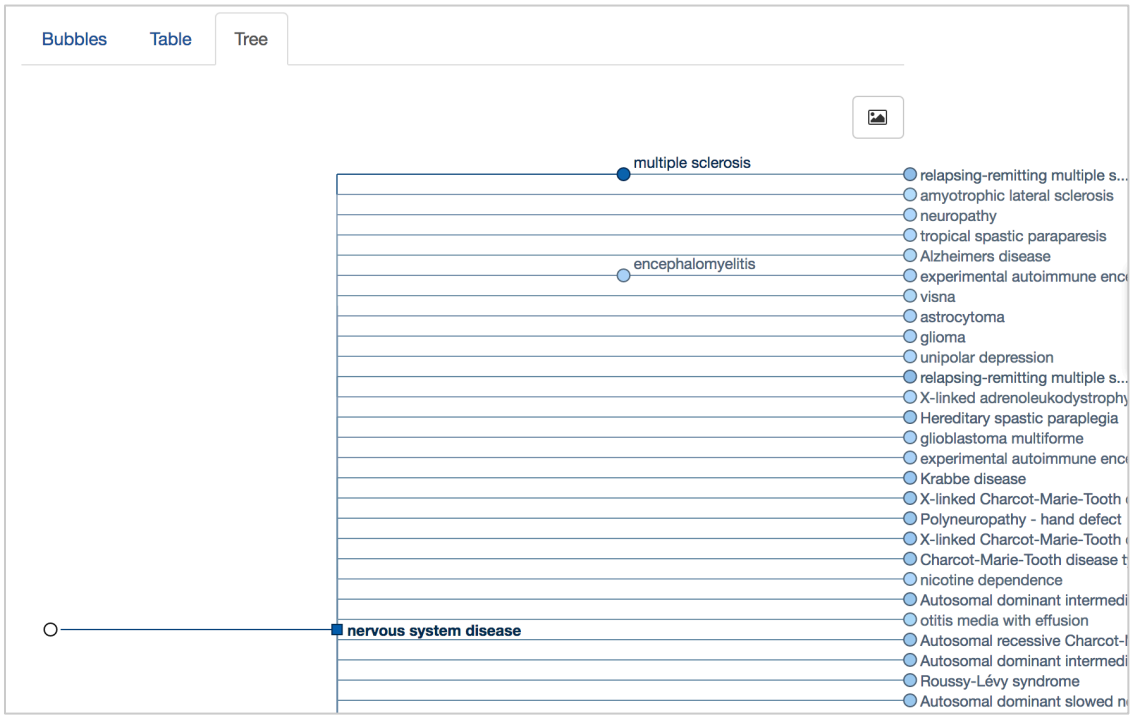
In this view, we list all diseases associated with a target, ordered by the association score, which is colour coded. When there is no evidence to support the association, the cells in this table are coloured in white (score of zero). You can show the 10 first entries and get the pagination for the remaining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

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

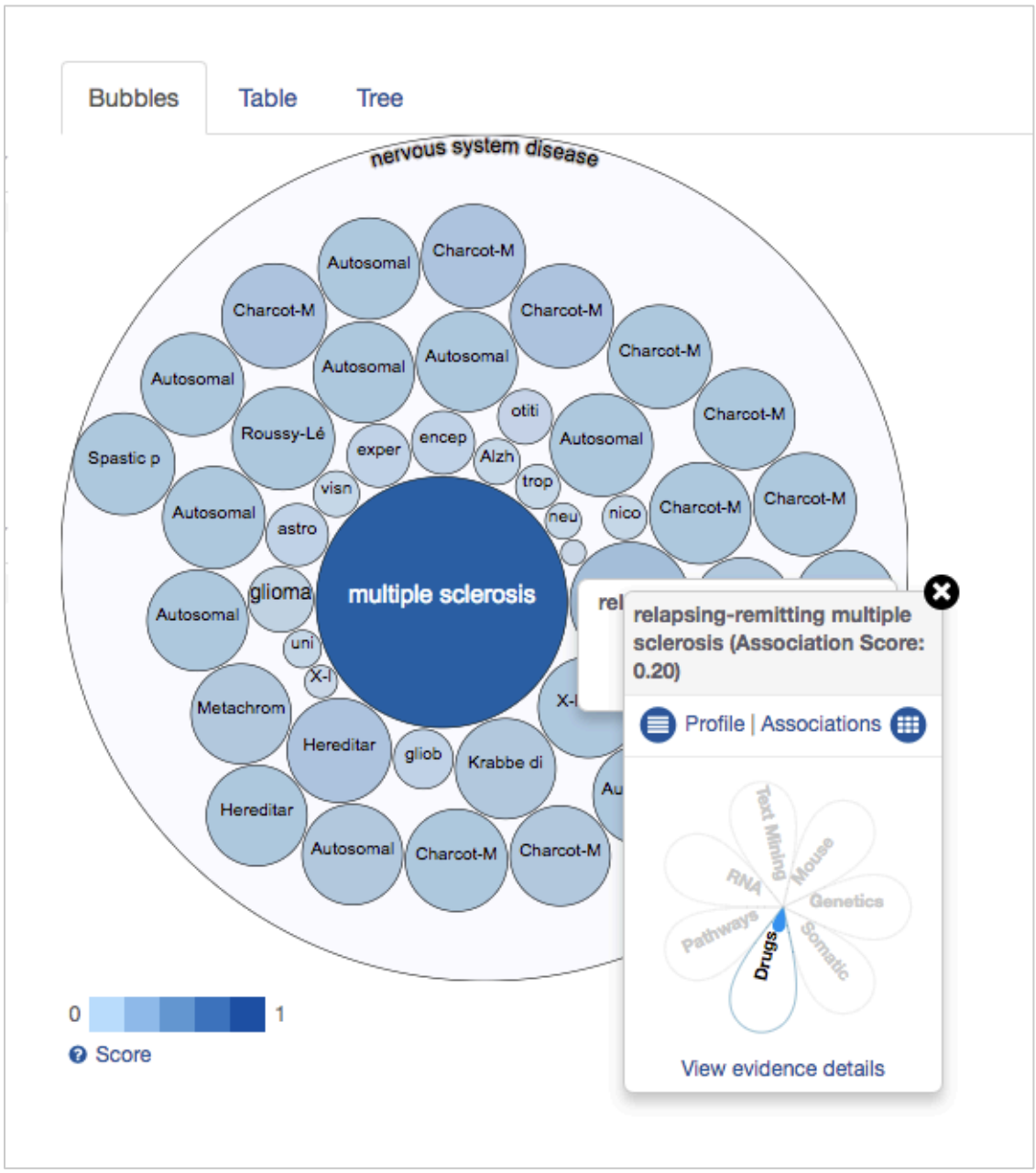
Tip: We colour code the cells in the table in different shade of blue as a visual way to convey the strength of the association (strongest association is coloured in dark blue). The score varies from 0 to 1. Hover over the cells in the table to view the numbers. Alternatively, you can select the cells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

Exercise 1

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) Can you use the platform to find out if this drug is currently used in trials with other dementia related diseases?

b) Which genes does this drug match to?

c) Can you list the diseases from the nervous system where there is a strong confidence (>90%) in the association with this gene?

Let's focus now on the different evidence used to associate *APP* with Alzheimer's.

d) Are there any known genetic variants (i.e. with a reference ID such as rs123456) listed in the Genetic associations table? Can you find some of the papers that support this association?

e) Click on the 'Browser' link to view the mutations is a graphical display. Are there variants associated with other traits (or diseases) in the region of the *APP* gene?

f) Which biochemical pathways seem to be affected by pathogenic mutations in this target?

Let's now have a look at the target itself by clicking on the target name. This page provides a profile for a target and it's where you can explore more information on a target such as data on RNA baseline expression, gene tree with orthologous genes in other species, etc.

g) Which tissues according to the GTEx project does this target seem to be highly expressed?

Exercise 2

LRRK2 in Parkinson's disease

BACKGROUND
The *LRRK2* gene encodes a protein with 5 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) Is this target associated with other diseases of the nervous system? Which type of evidence seems to suggest such associations?

b) How long is the protein encoded by this gene/target? Can you find the domains listed above in the BACKGROUND section?

c) Can you use the Open Targets Platform to find which chromosome this gene maps to on the human genome? Is it a forward or reverse stranded gene?

d) Which tissue has the highest RNA baseline expression according to the GTEx project?

Exercise 3

Filter the target association table for Alzheimer's based on a list of known targets.

This is a list of several genes that seem to be associated with Alzheimer's based literature searches:

HFE
PSEN1
TF
APOE
ADRB2
PSEN2
A2M

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

b) Are there any targets, which are membrane receptors?

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

Exercise 4

How can I retrieve all disease associations for several genes of interest, all at once?

BACKGROUND

So far you have used the website www.targetvalidation.org to search for target-disease associations on a gene by gene (or disease by disease) basis. You may want to access and retrieve data on several genes or several diseases. For this, you can either use the batch search tool for up to 200 targets. For a list of diseases (or a list of more than 200 targets), you can use our REST API (with or without our Python and R clients).

USE CASE
The following three genes have been associated with gastric carcinoma:
ENSG00000141736
ENSG00000141510
ENSG00000132356

QUESTIONS
a) "How can I find out all diseases (besides gastric carcinoma) associated with those three Ensembl gene IDs?"

b) "Which diseases have got the highest overall association score for each of those three genes?"

c) Can I download the above list in TAB format?"

Interested in other use cases using our REST API? Check our blog posts.

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 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 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 (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/

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/