Mining gene-disease associations for drug identification and discovery with Open Targets

Hands-on Workshop
Course booklet

CRUK Manchester Institute
22nd November 2016

Denise Carvalho-Silva
Open Targets Outreach

Notes

This workshop is based on v1.2.1 of the Open Targets Platform (September 2016 release)

Some useful links:

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

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

3) Workshop materials (in pdf)
https://github.com/deniseOme/training

4) Feedback survey
http://tinyurl.com/manc-221116

Feel free to tackle questions relative to your own research instead of following the ones provided in this course booklet. The answers for exercises 1 and 2 can be found here:
https://github.com/deniseOme/training

Questions or Feedback?

support@opentargets.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 validation of targets relevant to key therapeutic areas namely:

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

Finally, 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 release v1.2.1 (September 2016), the platform serves information on 30,591 targets; 9,425 diseases; 4.8 million evidence; and 2.4 million target-disease associations.

In addition to the web application, we include the data dumps and an API.

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 for this association
- 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.* (accepted)
Nucleic Acids Res (2017 Database Issue)

OPEN TARGETS PLATFORM: WALKTHROUGH

You have now had a chance to explore the Open Targets website (opentargets.org) and found out:

- More about the Open Targets consortium, including its core principles
- The types of cancer experimentally studied in the lab by members of the consortium
- The key challenge of the Core Bioinformatics team
- How to get to the Open Targets Platform

Let's now focus on the Open Targets Platform.

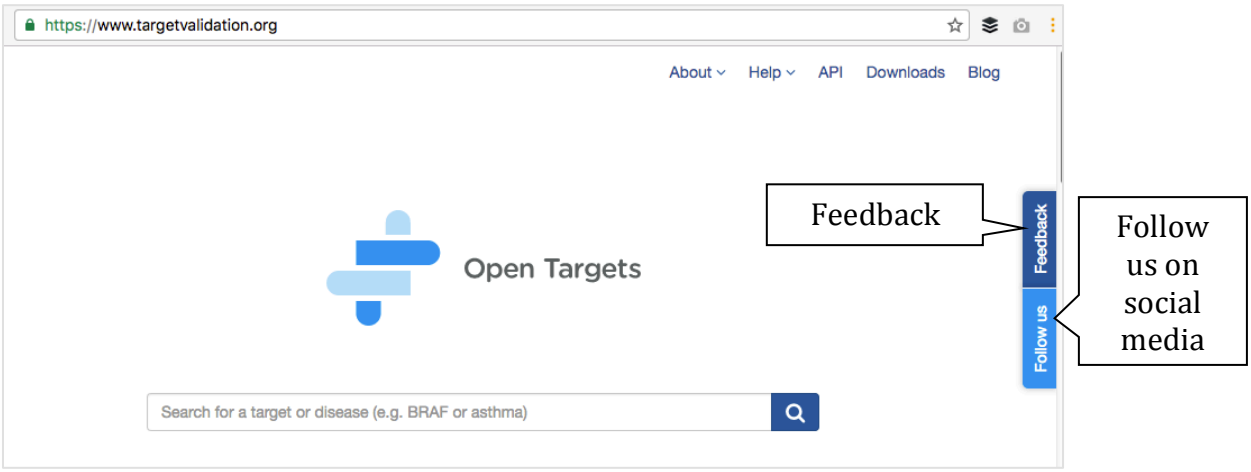
We will guide you through the website using congenital heart disease (CHD) as an example.

Sinfrim *et al* (Nature Genet 2016) has looked at the role for *de novo* mutations and mutations of incomplete penetrance in genes associated with CHD. The paper is entitled "Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing".

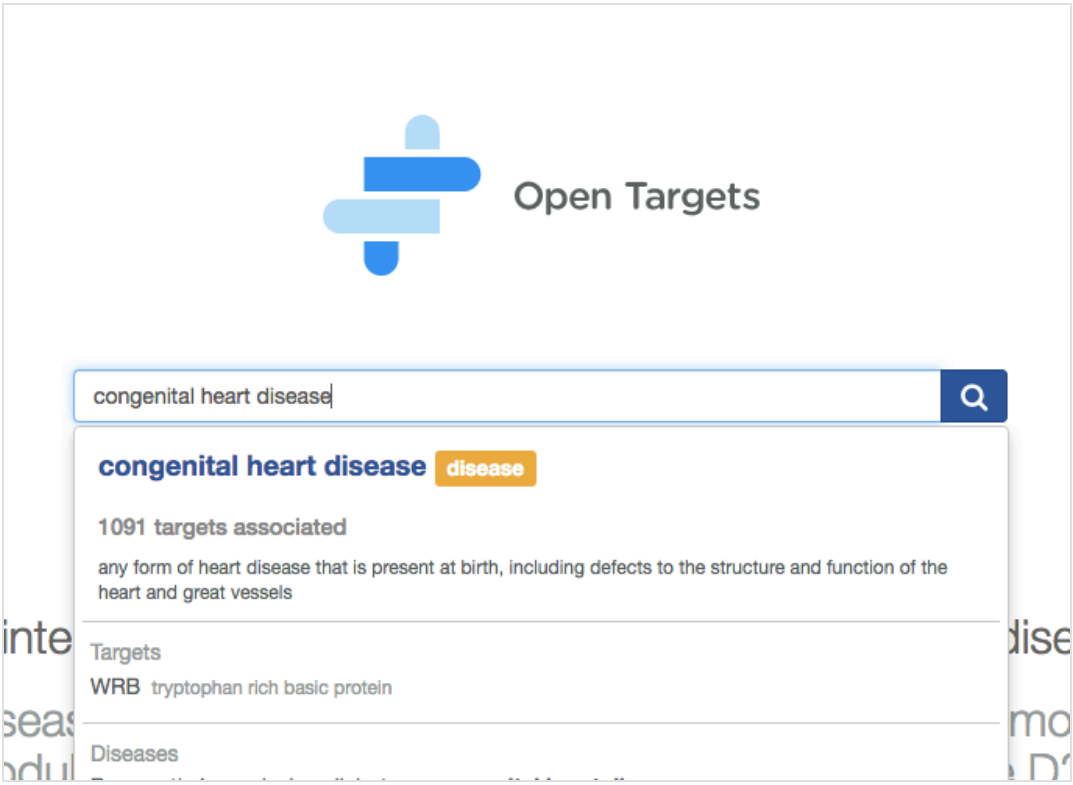
The following points will be addressed during the walkthrough:

- Targets associated with CHD
- Filter down the number of targets based on specific evidence
- Data sources used to support the target-CHD association
- Looking for other diseases associated than CHD with a target
- Visualise a target in a browser like view
- Find out how strong is the association between a target and a disease
- Find drugs currently in clinical trials

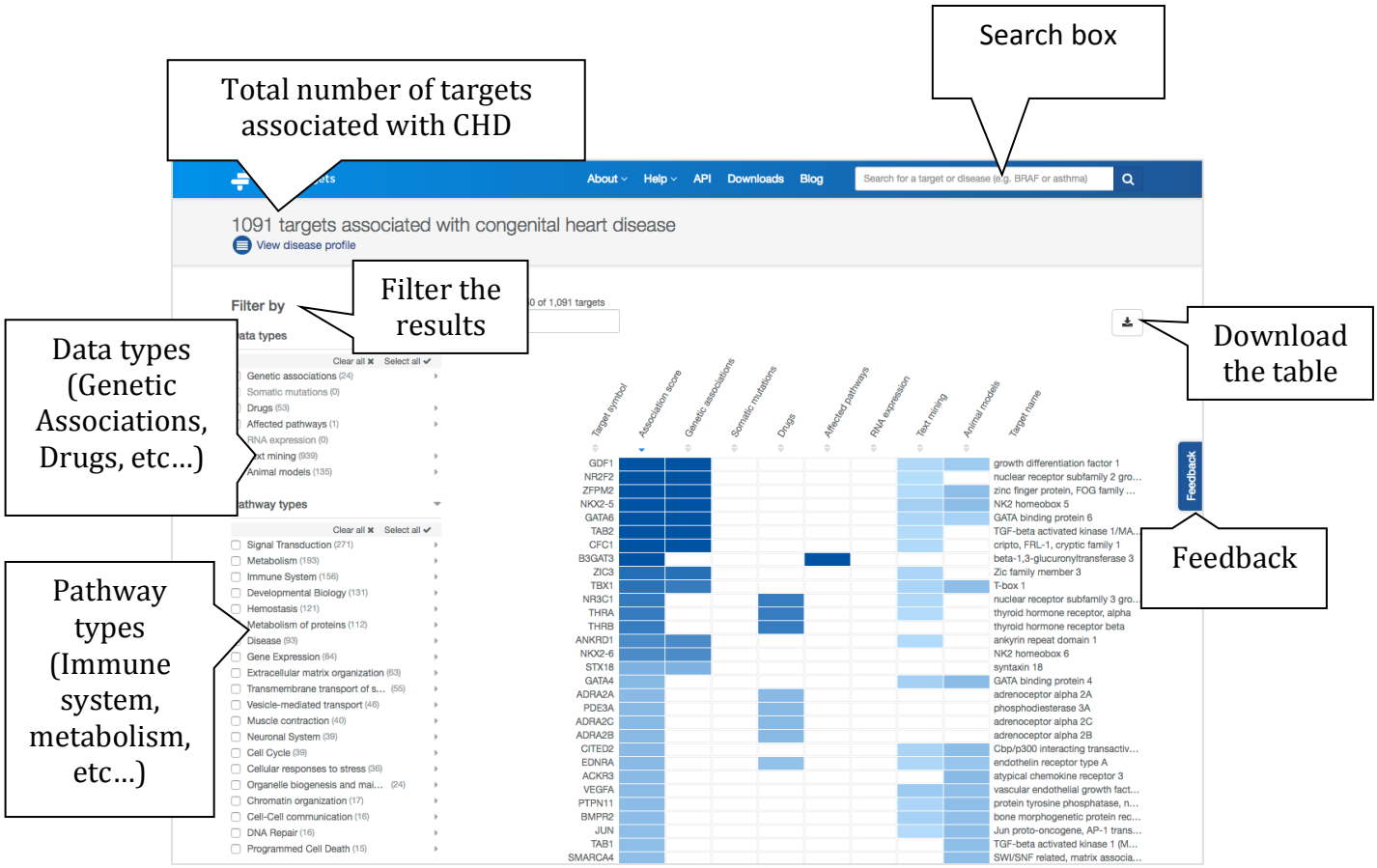
Go to www.targetvalidation.org and type in 'Congenital heart disease' in the search box below:



Select the first (best) hit:



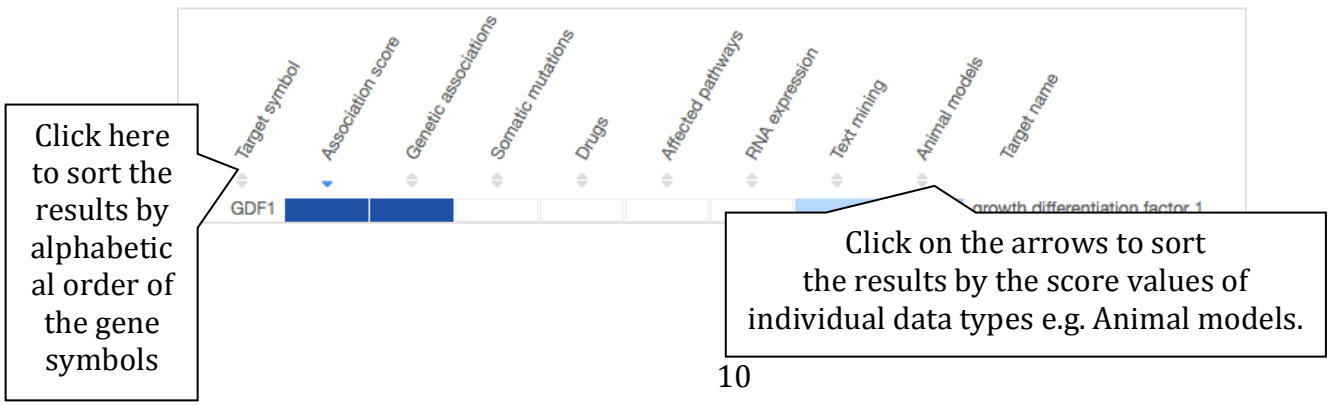
You will see a page like this:



The table is sorted by default with the best hit according to the highest association score on the top of the table.

You can sort the table listing the targets associated with the disease.

The sorting can be done by alphabetical order of the list of targets or numerically according to the association score (either overall) or the individual score for each 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 calculated for each piece of evidence that is used to support the association and the individual scores are combined to give the overall score ('Association score' column in the table below):



The current release of the Open Targets Platform (September 2016) lists 1,091 targets associated with CHD. You can filter this number according to 'Data types' or 'Pathway types':

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
- …

What are **Data types** and **Pathway 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.

When searching for a disease, we will list all the targets associated with it. You can filter the results and focus on targets seen in certain Pathway e.g. Signal Transduction, Cell Cycle, Immune System and much more.

Let's filter the data by 'Genetic associations' to show the targets associated with CHD, which are based on genetic variants (such as SNPs from GWAS or UniProt, for example) only. The number of targets goes down to 24.

The data sources used to support the association of these targets to CHD from the genetic point of view are:

GWAS catalog
UniProt
UniProt literature
European Variation Archive
Gene2Phenotype

Check our help pages on the different sources used in the Open Targets Platform:

https://targetvalidation.org/data_sources

If you wish to suggest data or resources we could incorporate in our platform, please email them to us (support@targetvalidation.org).

Let's now focus on one of these 24 targets, *GDF1*. Click on any cells in the *GDF1* row to get to a page containing all the 'Evidence for GDF-1 in congenital heart disease' i.e.

targetvalidation.org/evidence/ENSG00000130283/EFO_0005207

In this page, you can explore the details of the data types (e.g. Genetic association, Somatic mutations, Drugs, Text mining) that support the association between a target and a disease. These are shown in different tabs.

Note: if there is no data for a given data type of evidence, the tab will be grey out.

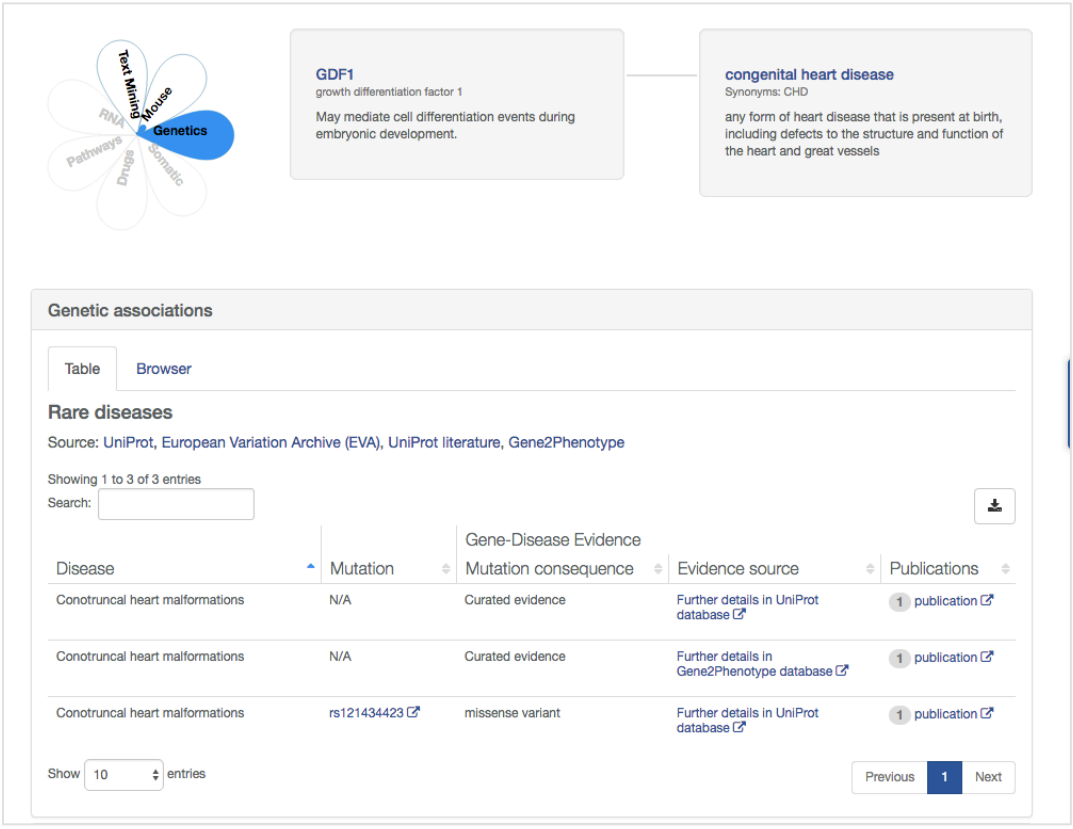
Alternatively click on the cell in column 'Genetic associations' to land in the page below:

https://targetvalidation.org/evidence/ENSG00000130283/EFO_0005207?view=sec:genetic_associations

where the tab for Genetic associations is already open.

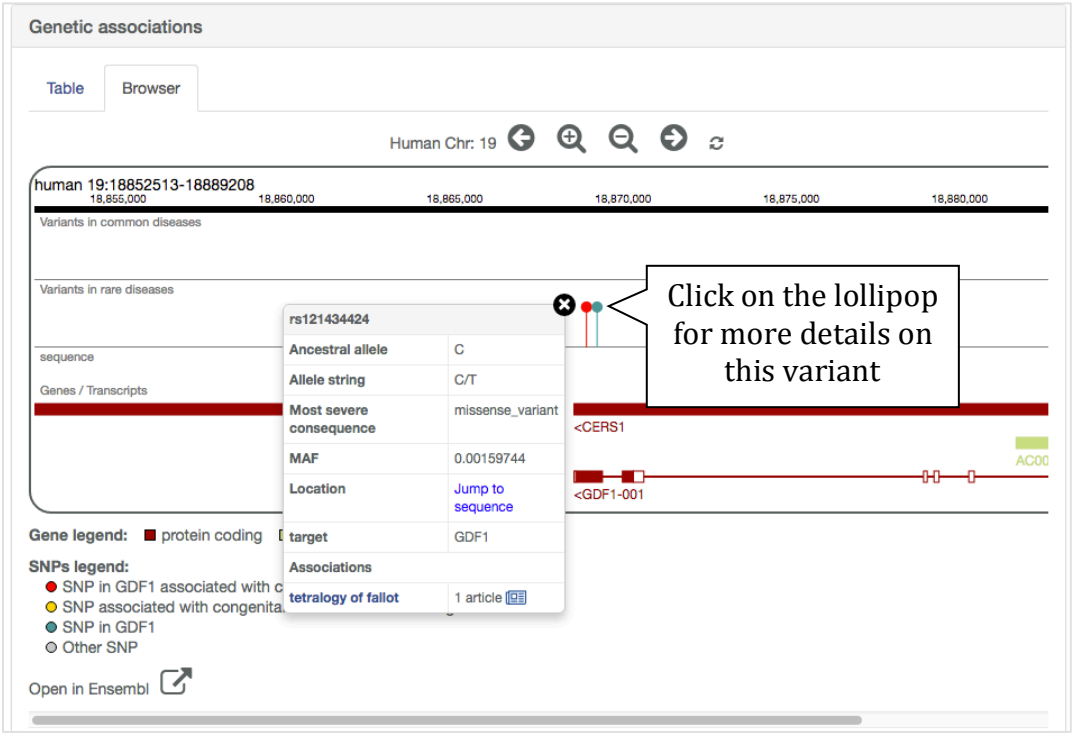
In the Genetic association table, you will get the genetic variants, the functional consequence (from Sequence Ontology) of the variants on the gene of interest, the source of the data (e.g. UniProt, Gene2Phenotype) and relevant publications.

Both targets and diseases are highlighted if they co-occur in the same sentence. The co-occurrence is automatically searched for in all papers from Europe PMC, including the title, abstracts, text, background and other sections of the article with the exception of supplementary tables:

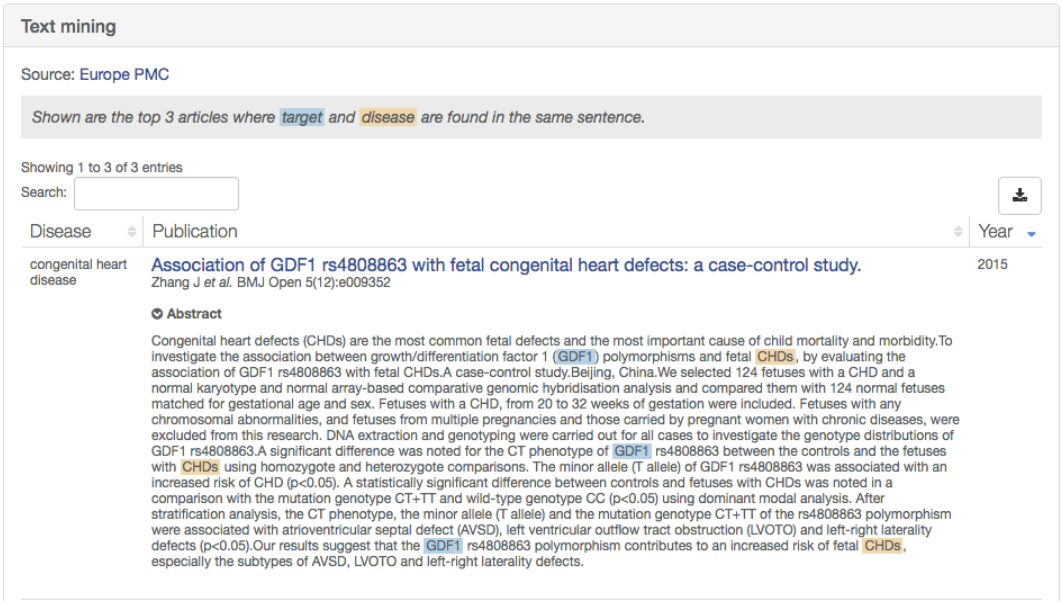


One of these variants is known in public databases and identified as rs121434423.

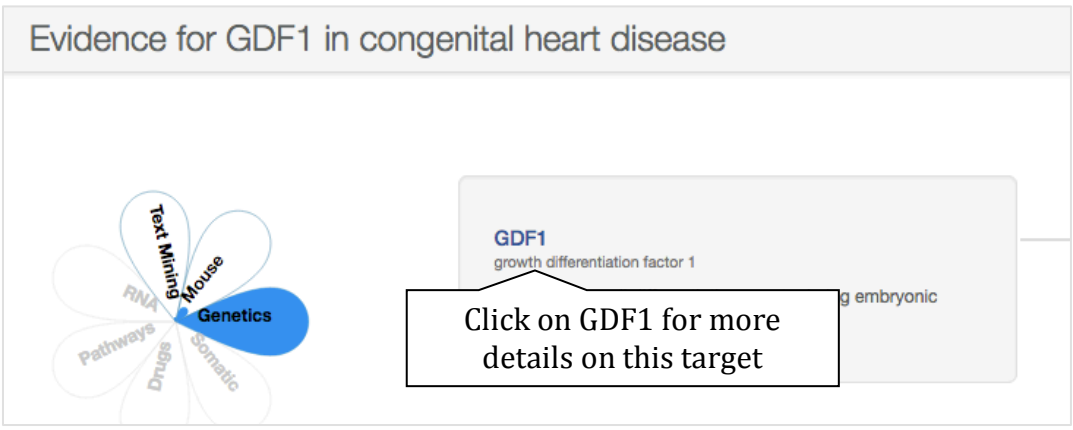
You can also explore this information in an interactive Browser view to zoom in and out, scroll along the genome and find out more about the gene (s), transcript (s), genetic variants (represented as lollipops). We also provide links to Ensembl:



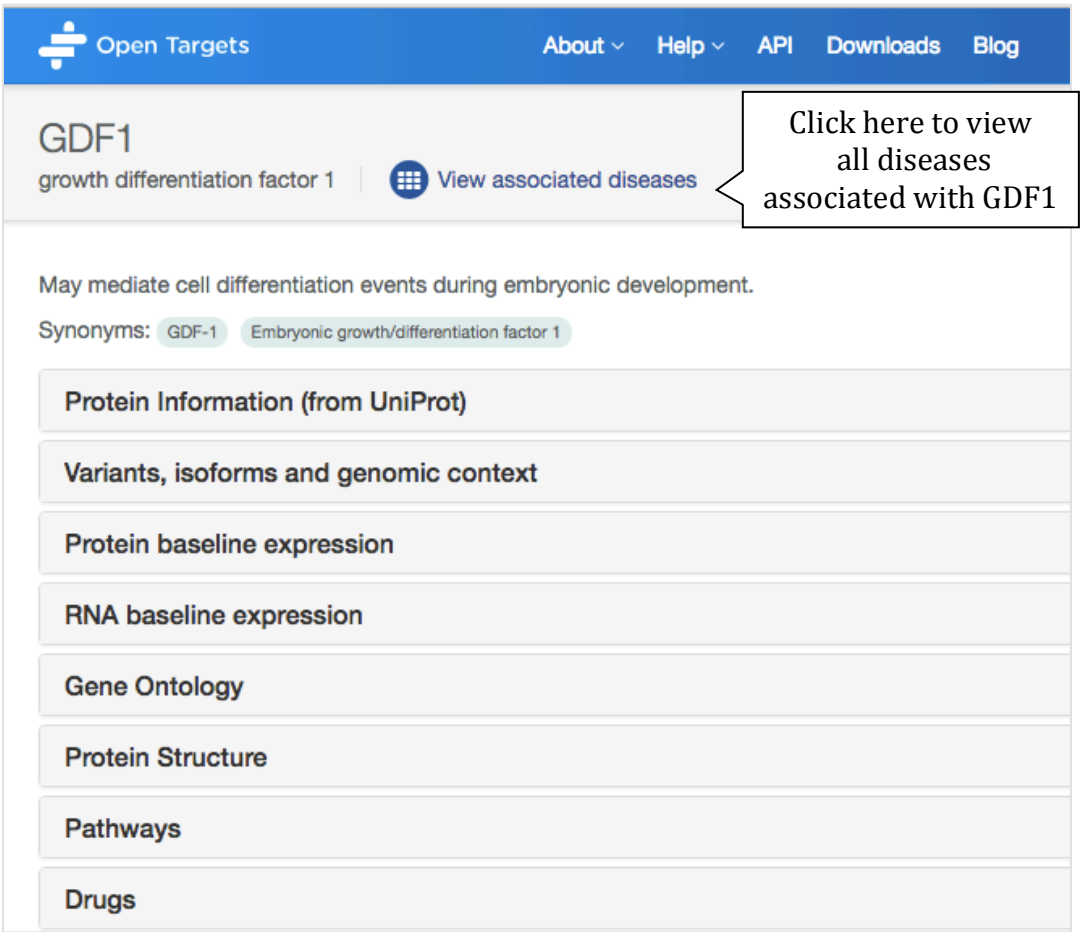
Let's now explore other pieces of evidence that seem to support the association between GDF1 with CHD. Still on the same page, scroll down to view the 'Text mining' tab. Click on it to see the three papers support the association:



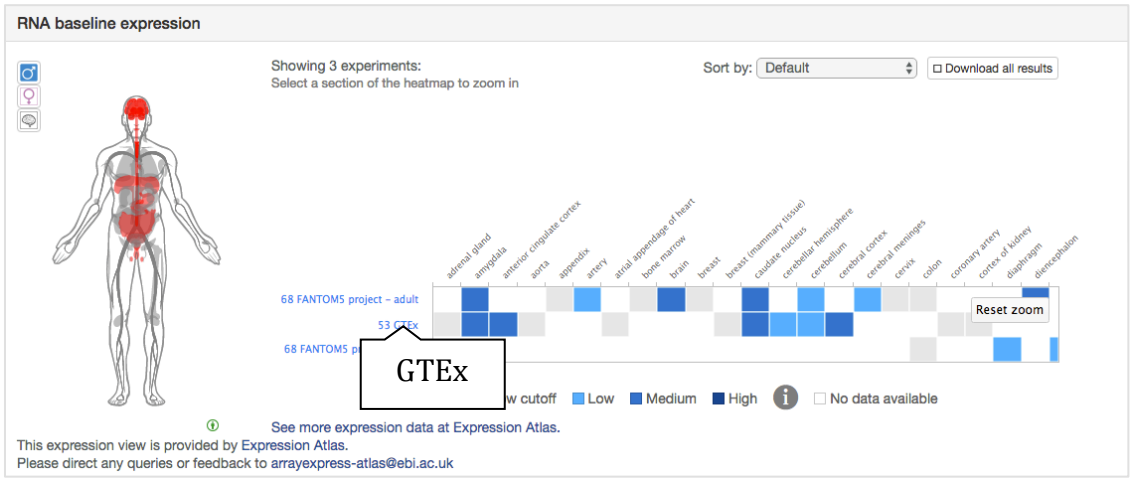
We can now explore the target in more detail and find out its baseline expression (at the RNA level), for example. On the same page, scroll back up and click on the GDF1 link next to the flower:



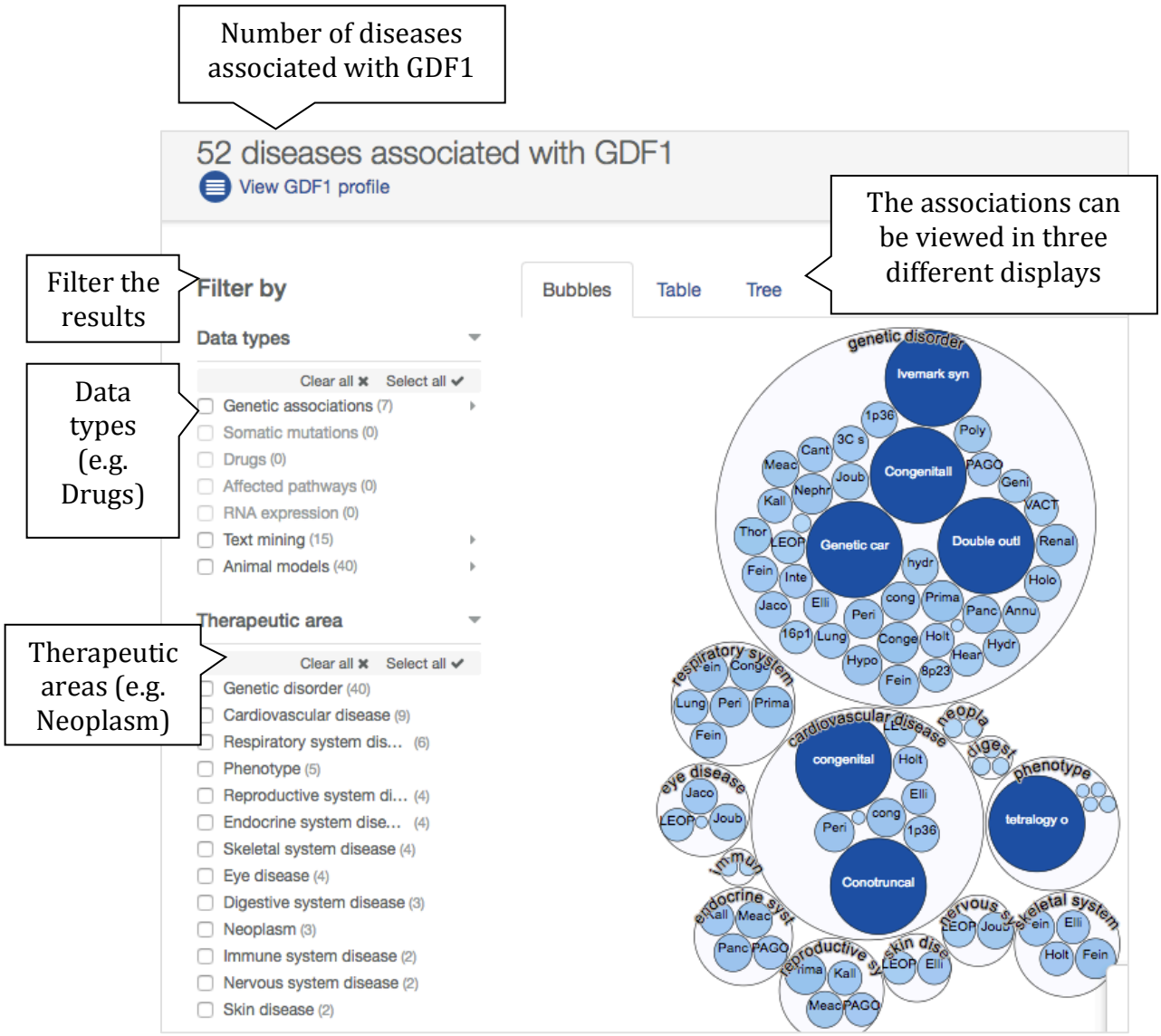
You will land on a page that gives you target specific details such as Protein information from UniProt, Variants, isoforms and genomic context (from Ensembl), Gene Ontology, Pathways (from Reactome) and much more:



Expand the RNA baseline expression to find out which tissues seem to be highly expressed (coloured in dark blue) according to the GTEx project, such as hippocampus and cerebral cortex:



Let's now click on 'View associated diseases' to find out all diseases associated with the GDF1. You will land on a page like this:



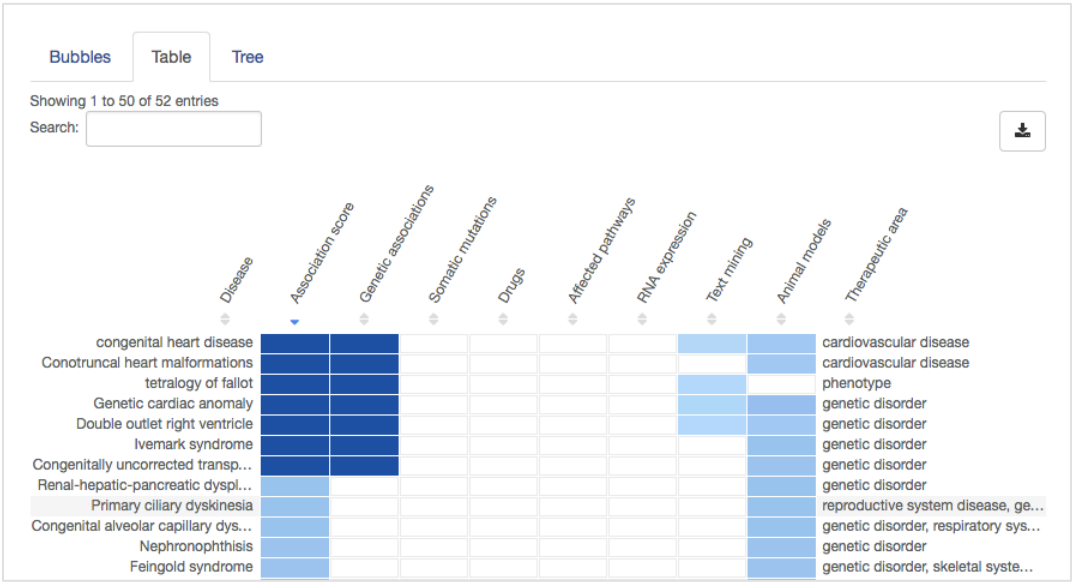
There are three different displays to view the diseases associated with a target:

- 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.

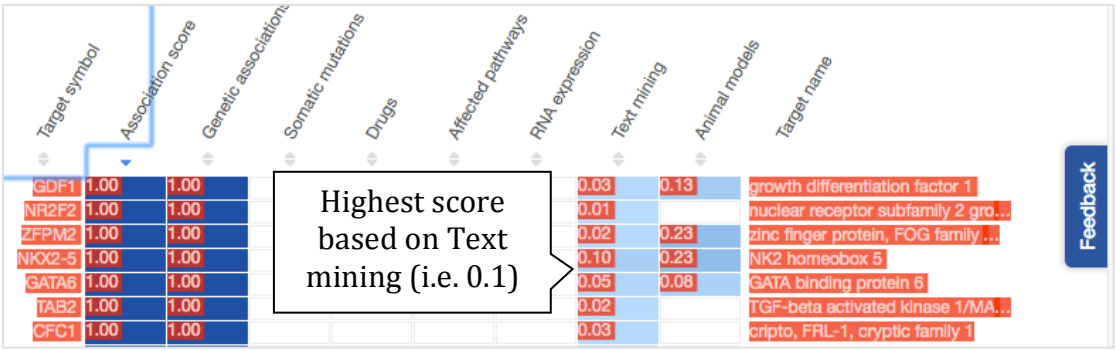
- Table view

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):



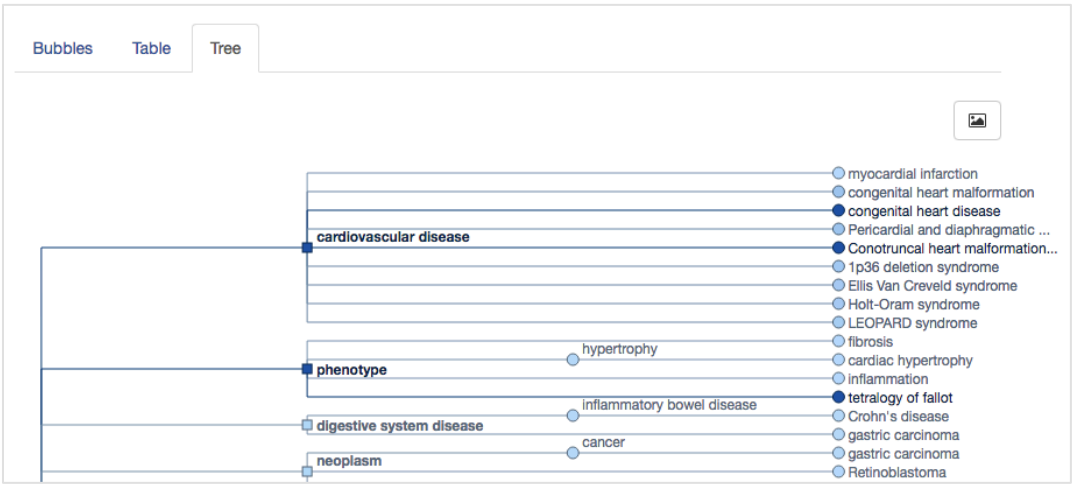
Evidence from highly specific terms of the disease ontology is aggregated to broader, parent terms. You can order the associations by their scores for individual data types (e.g. Genetic associations, Somatic mutations).

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. In order to view the actual number, you hover over the cell. You can also select the cells in the table so that you can view the numerical values.



- Tree view

In the Tree view, you 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:



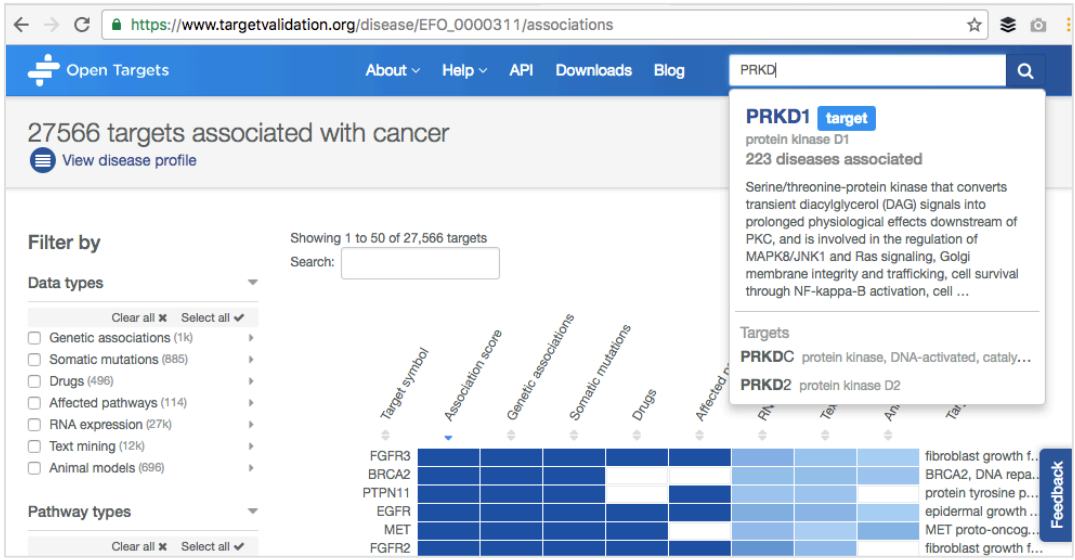
Regardless the view you choose to explore, you can filter the data to find out if there are other cardiovascular diseases associated with this target. There are eight of them (plus coronary heart disease, CHD) including conotruncal heart malformations (with the highest score of 1) plus a few others with score ranging from 0.17-0.12 (e.g. LEOPARD

syndrome) and one disease with score of 0.02 (i.e. myocardial infarction).

So far, we have explored the Open Targets Platform by starting off with a disease (i.e. congenital heart disease, or CHD). You can also search for genes (soon we will index drugs and genetic variants such as SNPs and searching those will be available as well. So stay tuned).

Let's have a look at the *PRKD1* gene, one of the targets described in the paper "Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing" by by Sinfrim *et al* (Nature Genet 2016).

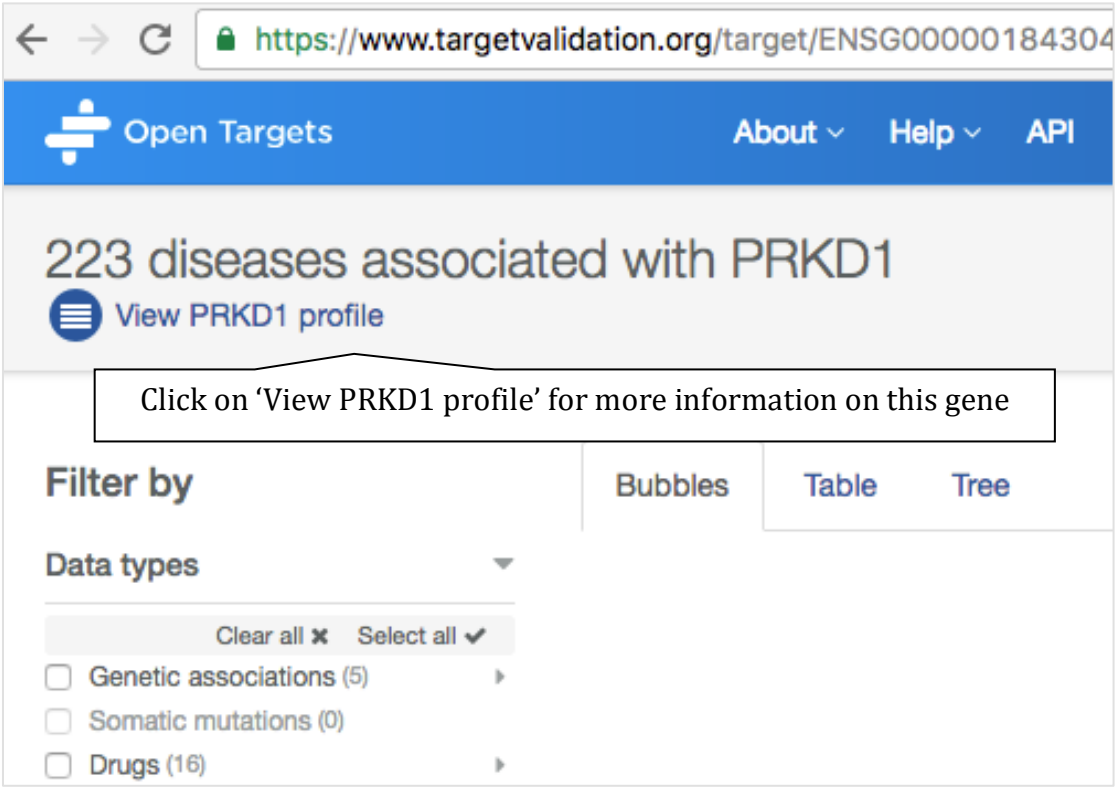
Search for that target using the search box at the right corner of any page in our Platform:



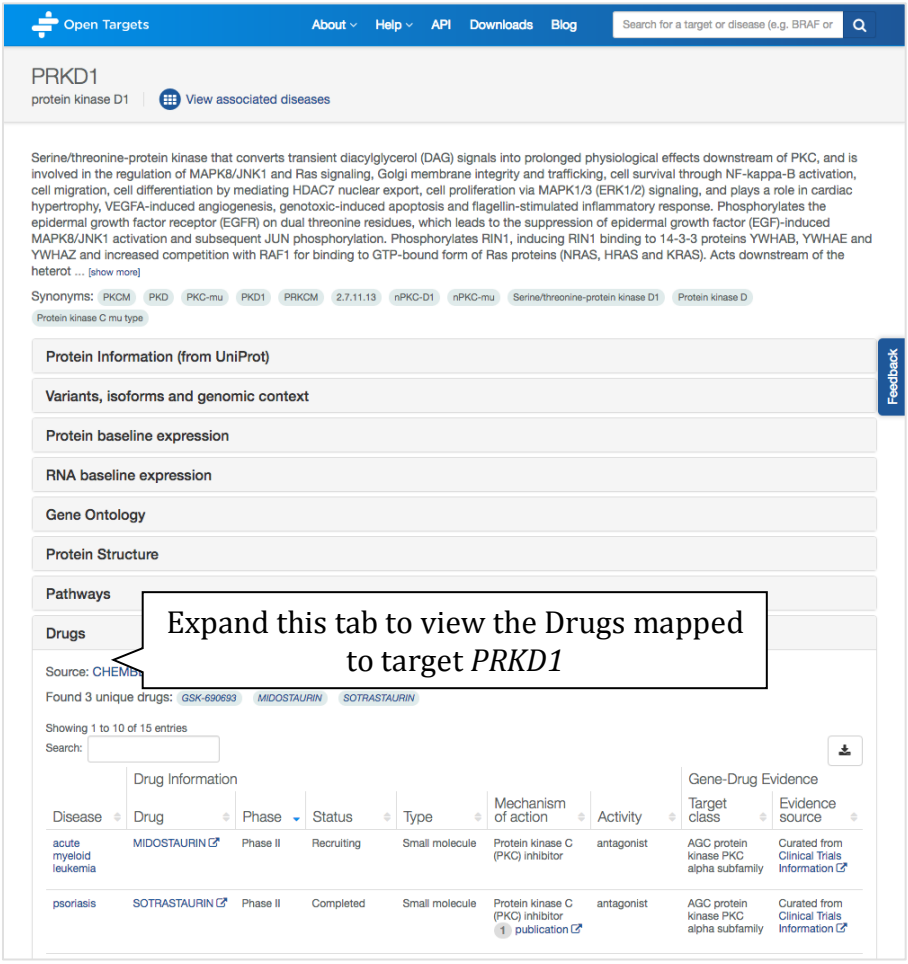
Alternatively, you can go back to the homepage and search for that target from the search box in there:

www.targetvalidation.org

Either way you will land on a page showing the diseases associated with *PRKD1*.



Let's now click on the 'View PTPN11 profile'. Then on 'Drugs' to find out which ones are currently in clinical trials and could be affect this gene:



The source for this data is from ChEMBL. There are three unique drugs mapping to this target, i.e. SOTRASTAURIN, MIDOSTAURIN and GSK-690693.

They are in different phases of clinical trials (such as phase II or I) and under investigation for different cancer types such as acute myeloid leukemia and mast-cell leukemia.

END OF THE WALKTHROUGH

HANDS-ON EXERCISES

Exercise 1 – Prioritising targets for drug discovery in prostate carcinoma

BACKGROUND
Prostate carcinoma is the most common type of cancer in men in the UK. More than 41,000 cases are newly diagnosed every year. The causes of prostate carcinoma are unknown. Age, ethnic background and family history are some of the factors that can increase one's risk of developing the condition (source: NHS choices; Cancer Research UK).

SIGNIFICANCE
Men with a father or brother diagnosed with prostate carcinoma are two to three times more likely to get the condition, compared to the average man. The risk of developing this type of cancer is also higher risk of prostate carcinoma if their mother has had breast cancer. Some of the genes that seem to be associated with prostate carcinoma are *BRCA1* and *BRCA2*.

QUESTIONS
a) What are the top 10 targets associated with this prostate carcinoma when taking into account all pieces of evidence integrated in the Platform?

b) Restrict the search based on targets for which the association with the disease was based on Somatic mutations. Is this refined list the same as the one resulting from step (a) above?

Let's focus on one of these targets namely *FGFR4* and find out more about some of the evidence that seems to support the association.

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

d) Can you view these mutations is a graphical display? Are there other variants associated with other traits (or diseases) in the region of the *FGFR4* gene?

Let's now have a look at the target itself and explore more information on the *FGFR4* gene such as the data on RNA baseline expression.

e) What is the tissue with the highest RNA baseline expression level according to Human Proteome Map (in adult tissues)?

f) Have a look at the graphical view of the Protein information (from UniProt) and explore the Topology information. Which amino acids correspond to the transmembrane (TM) domain? Did you expect this protein to have TM domains? Why?

Exercise 2 – *MS4A1* as a possible drug target in the treatment of non-Hodgkin's lymphoma

BACKGROUND
The B-lymphocyte antigen CD20 is an activated-glycosylated phosphoprotein expressed on the surface of all B-cells beginning at the pro-B phase and progressively increasing in concentration until maturity. In humans, CD20 is encoded by the *MS4A1* gene.

SIGNIFICANCE
CD20 is the target of monoclonal antibodies (mAb) in the treatment of all B cell lymphomas, leukemias and autoimmune diseases. Some of these active agents (mAb) are on clinical trials for non-Hodgkin's lymphoma. Others anti-CD20 mAB have been approved by the FDA for B-cell chronic lymphocytic leukemia.

QUESTIONS
a) How many diseases within the broader therapeutic area of 'Hematological system' are associated with this target based on 'Drugs' data types as evidence?

b) Can you get a table with the diseases associated with this target? Can you name a few diseases with an association score equal or above 90%? Which format can you download this table as?

c) In addition to the data evidence ‘Drugs’, are there other types of evidence supporting the association between *MS4A1* and non-Hodgkin's lymphoma?

d) What are the drugs used as evidence to associate *MS4A1* with non-Hodgkin's lymphoma? Which drugs are in clinical phase III of their trials?

e) Can you find all drugs linked to this disease (therefore targeting other genes, not only *MS4A1*)?

f) Can you find the ontology of non-Hodgkin's lymphoma and name the different subtypes (children terms of non-Hodgkin's lymphoma?). Can you download this image? Note: you may want to click on the children diseases and see which targets have been associated with them.

EXTRA HANDS-ON EXERCISES

If you have finished exercises 1 and 2 above, you may want to do these extra ones:

Exercise 3 – The IKBKE gene, a noncanonical I-kappa-B kinase

BACKGROUND
The IKBKE gene has been identified as a breast cancer oncogene and is amplified and overexpressed in over 30% of breast carcinomas and breast cancer cell lines.

QUESTIONS
a) How many amino acids does the gene encode for?

b) Which amino acids correspond to the kinase domain? Are the other domains mapped to this protein?

c) What is the RNA baseline expression of this gene in breast tissues?

d) Can you list a few examples of biological processes the IKBKE protein may be involved with according to the Gene Ontology consortium?

e) You may want to use a mouse model to perform functional studies of IKBKE. But you need to find out whether or not an orthologue of IKBKE in mouse has been annotated. Can you find a gene tree for the IKBKE gene? Is there a IKBKE gene in mouse?

f) Which diseases seem to be associated with IKBKE in neoplasm? Are there somatic mutations supporting any of these associations?

Exercise 4 – Non-small cell lung carcinoma and possible drug targets

BACKGROUND

Non small cell lung cancer (NSCLC) is the most common type of lung cancer. There are three common types of non small cell lung cancer, which make up about 87 out of 100 lung cancers in the UK.

QUESTIONS
a) Which target has the strongest association score of 1 for non-small cell lung carcinoma based on evidence for somatic mutations?

b) Can you find out whether those somatic mutations in this target are missense or other mutation type?

c) Are there animal models available to study this target in non small cell lung cancer?

Exercise 5 – Using the Open Targets Platform to find out if the modulation of a target by a drug poses any possible unsafe interactions or effects.

BACKGROUND
The main goals of drug development are effectiveness and safety. Although no drug is 100% safe (they all have side effects), the benefits of the drugs should outweigh the known risks.

SIGNIFICANCE
Many drugs used on the treatments of diseases can interfere with other physiological processes and even cause death when taken in excess. One of the ways to start assessing the safety of a new compound is to look at which target it modulates, whether or not this target is involved in other therapeutic areas such as cardiovascular and reproductive system, and the expression of the gene (or protein) in normal tissues.

USE CASE
Abemaciclib has been shown in vitro to be a selective ATP-competitive inhibitor of CDK6 kinase activity that prevents the phosphorylation and subsequent inactivation of the Rb tumor suppressor protein, thereby inducing G1 cell cycle arrest and inhibition of cell proliferation. Abemaciclib is being investigated in clinical trials in patients with breast carcinoma among other types of cancer.

QUESTIONS

a) Which data types support an association between CDK2 and breast cancer?

b) Are there other drugs in addition to abemaciclib used in clinical trials modulating the same target for breast carcinoma? Has any drugs been approved by the FDA?

c) What is the level of RNA baseline expression of the target (i.e. CDK2) in heart from the Human Protein Atlas study?

e) Is this target associated with cardiovascular diseases with a strong confidence (i.e. score of 1)?

Exercise 6 – How can I retrieve all disease associations for three 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 access our data in programmatic way.

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 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/