Mining gene-disease associations and drug target validation with Open Targets

Hands-on Workshop

University of Cambridge
17th October 2016

Denise Carvalho-Silva
Open Targets Outreach

Notes:

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

Some useful links:

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

2) About the Target Validation Platform
www.targetvalidation.org/about

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

4) Feedback survey
http://tinyurl.com/cam-171016

Feel free to tackle questions relative to your own research instead of following the ones provided in this course booklet. The answers for the latter 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 Target Validation (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 Target Validation Platform
- Browsing the Target Validation 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 validation data. The core also designed, created and maintains the Target Validation 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 the Target Validation Platform

The Target Validation 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 Target Validation 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 Target Validation 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 Target Validation 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.* (submitted)
Nucleic Acids Res (2017 Database Issue)

PLATFORM WALKTHROUGH

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 the disease. This mutation architecture could explain the low sibling recurrence risk observed in the clinic. 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:

- o Targets associated with CHD
- o Different likelihood of association between targets and CHD
- o Find out if there are animal models for a given target
- o Explore the sources for the evidence used
- o Looking for other diseases associated with a target
- o Having a closer look at one target

Go to www.targetvalidation.org and type in 'Congenital heart disease' in the search box below:
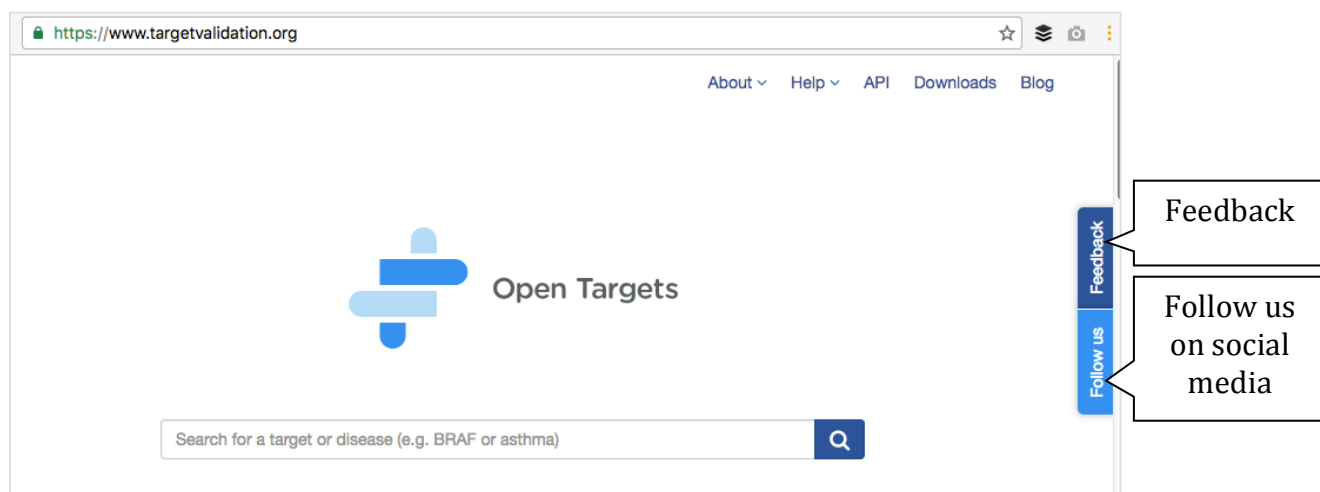
Select the first (best) hit:
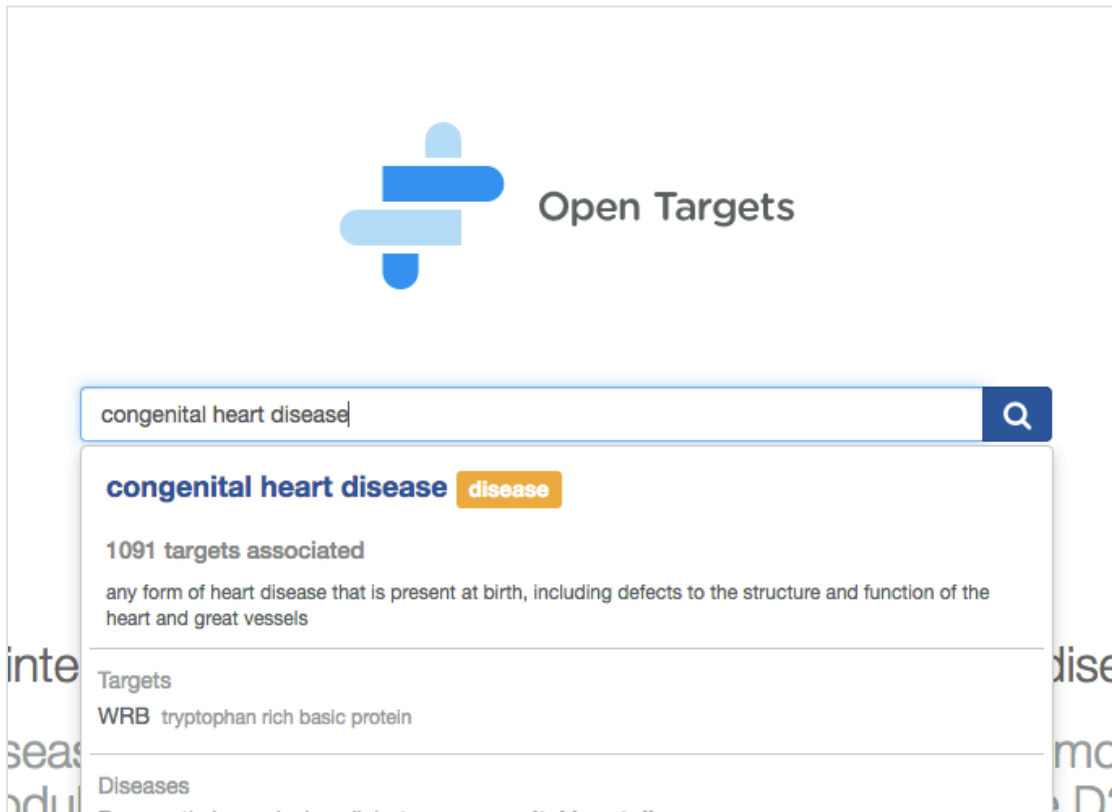
You will see a page like this:

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 piece of evidence (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 isused to support the association and the individual scores are combined to give the overall score (second column in the table below):

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

Click on the arrows to sort the results by the score values of individual data types e.g. Drugs, Text Mining.

The current release of the Target Validation Platform (version 1.2.1) lists 1,091 targets associated with CHD. You can filter this number according to either 'Data types' or 'Pathway types'.

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 to give rise to 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 search for a disease, we will list all the targets associated with it. You can filter the results and focus on Pathway types containing a target, 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 first seven rows contain the top seven targets with maximum association score of 1, namely *GDF1*, *NR2F2*, *ZFPM2*, *NKX2-5*, *GATA6*, *TAB2*, and *CFC1*.
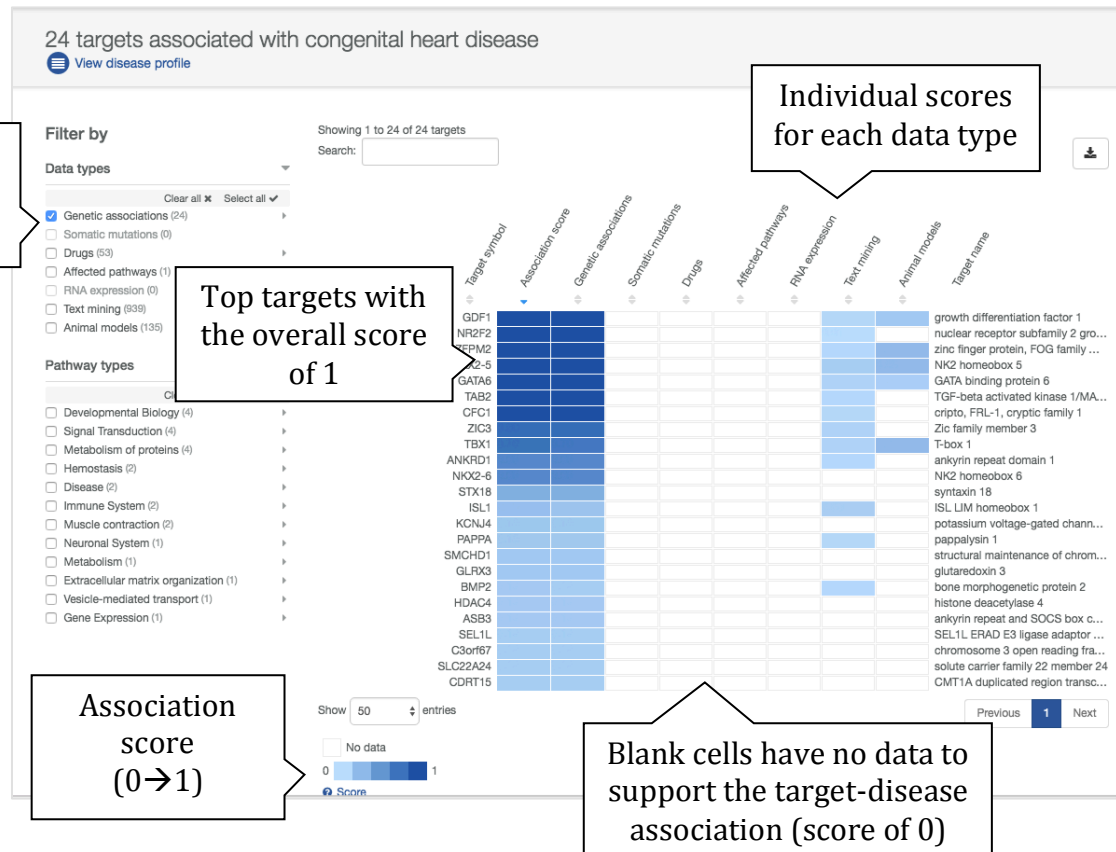
Among those, the target with the highest score based on Text mining is *NKX2-5.*

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.

Highest score based on Text mining (i.e. 0.1)

Let's now focus on *NKX2-5*. Click on any cells in the *NKX2-5* row to get to a page containing all the 'Evidence for NKX2-5 in congenital heart disease'.

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.

For the association between *NKX2-5* and CHD, we have data for the following data types:

- Genetic association: a table with the genetic variants (e.g. SNP ID), the functional consequence of the variants on the gene (from Sequence Ontology) of interest, the source of the data and relevant publications:

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:
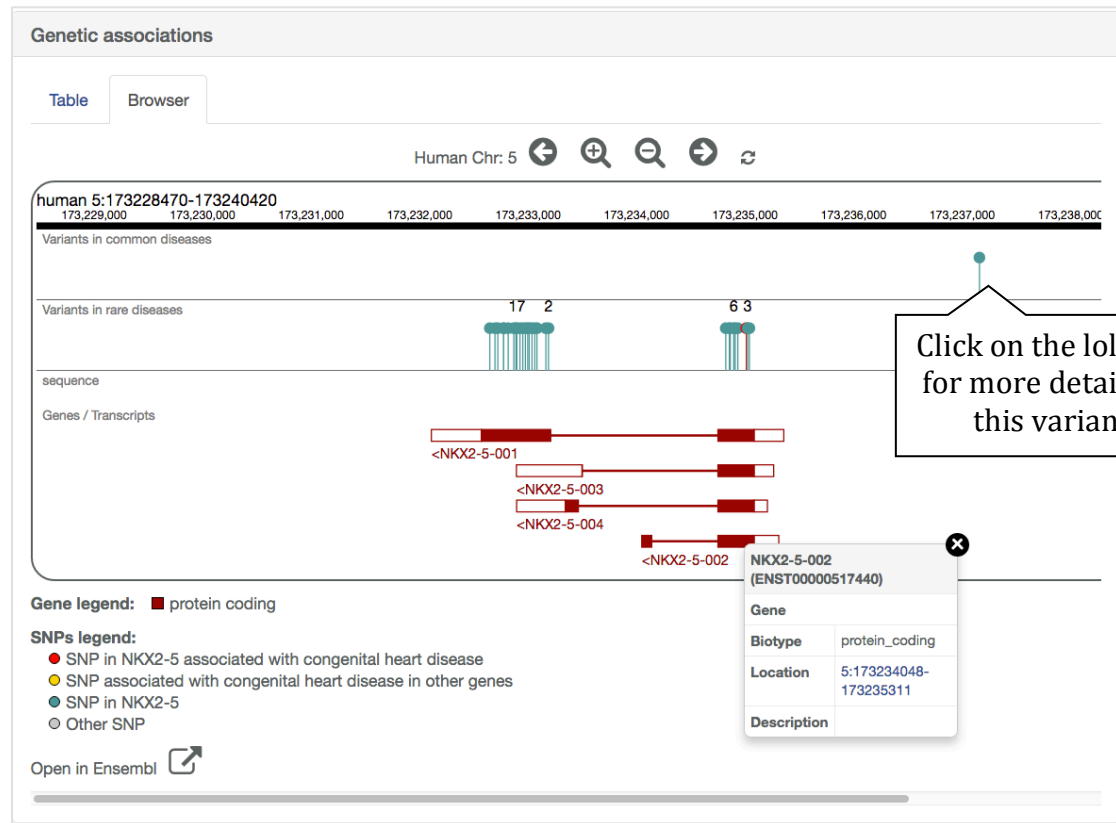
- Text mining: the papers that contain both target and disease in a sentence from the pipeline by Europe PMC. Both targets and diseases are highlighted. Both target and disease entities should occur in the same sentence. The co-occurrence is automatically searched for in all papers from Europe PMC, including the title, abstracts, text, background and other sections of the article with the exception of supplementary tables:

- Animal models: list of mouse models available where the target gene was disrupted (e.g. by knockout) and the resulting phenotype in mouse and the known phenotype in humans. In this section, you can get additional details on the mouse model if you click on the links to the MGI database:

The association between *NKX2-5* and CHD for the data type Animal models comes from the Phenodigm resource, based at the Wellcome Trust Sanger Institute.

Let's now find out if this target is associated with other diseases than CHD.

Scroll up to the top of the page till you see the flower and a link to the *NKX2-5*. Click on this link:
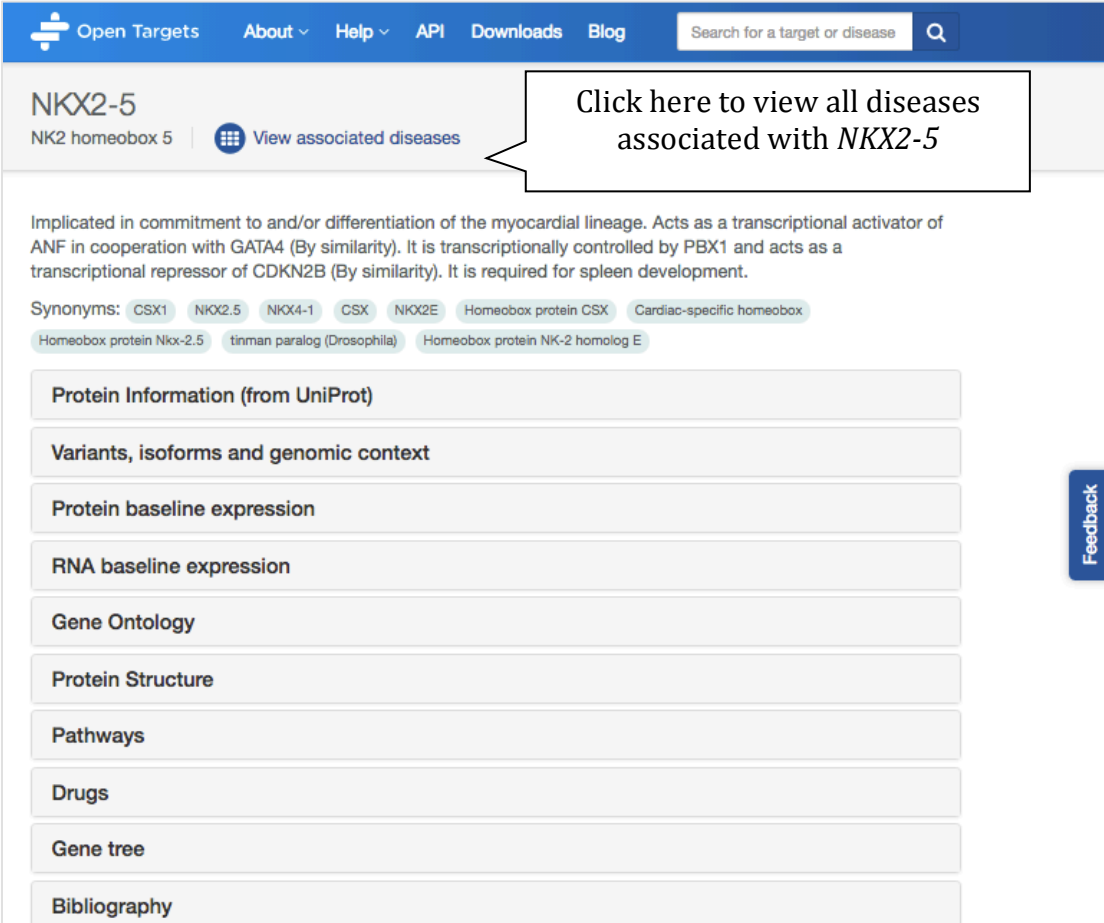
You will land on a page that gives you target specific details such as Protein information from UniProt, Variants, isoforms and genomic context (from Ensembl), Gene Ontology, Pathways (from Reactome) and much more:

Besides exploring the target (gene) information outside any disease context, from the page above (i.e. https://www.targetvalidation.org/target/ENSG00000183072), you can click on 'View associated diseases' to find out all diseases associated with the target *NKX2-5.* You will land on a page like this:
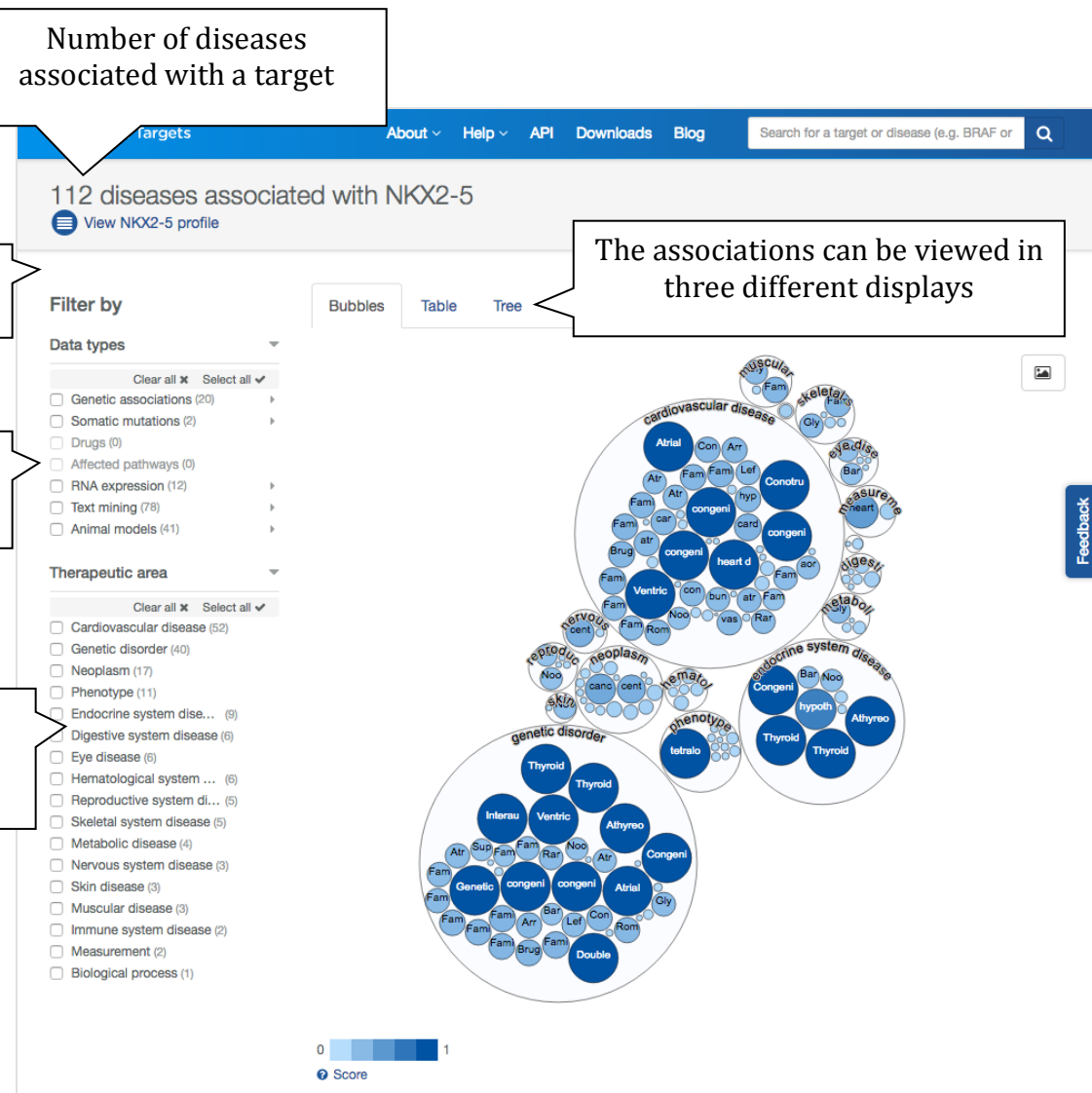
There are three different displays to view the data:

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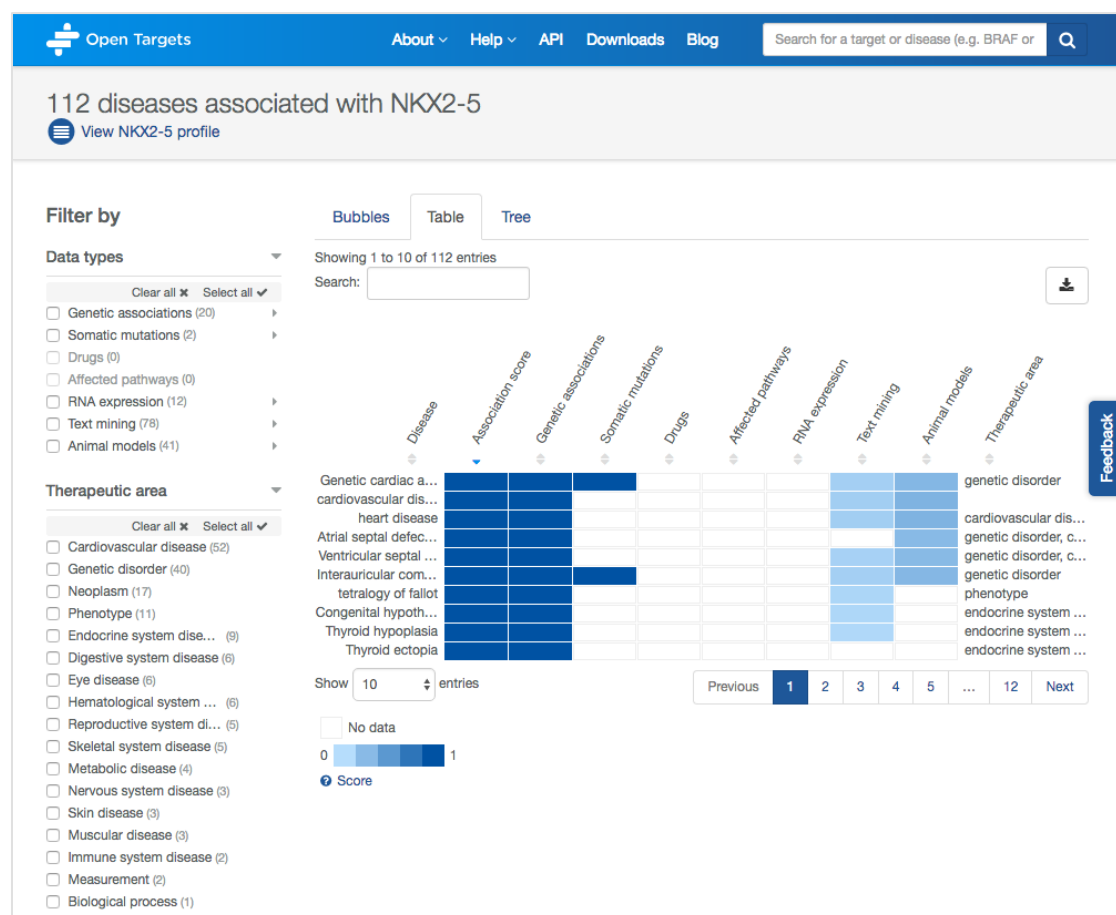
Evidence from highly specific terms of the disease ontology is aggregated to broader, parent terms. You can order the associations by their scores for individual data types (e.g. Genetic associations, Somatic mutations).

- 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.

In whichever view you are at, you can filter the data to find out if there are other cardiovascular diseases associated with this target. There are 52 of them, including Ventricular septal defect and aortic stenosis.
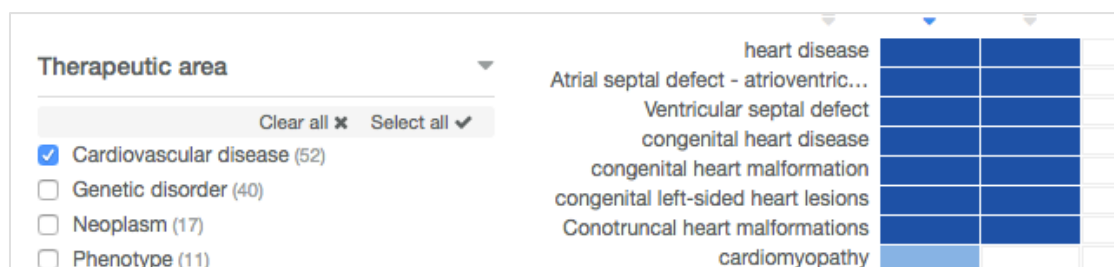
So far, we have explored the Target Validation Platform by starting off with a disease (i.e. Congenital Heart Disease, or CHD). You can also search for genes (soon we will index drugs and genetic variants such as SNPs and searching those will be available as well).

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).

Let's 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*.

Click on 'View PRKD1 profile' to get more detailed information on this gene

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 targeting this gene:

The source for this data is from ChEMBL. There are three unique drugs mapping to this target. They are in different phases of clinical trials (such as phase II or I) and under investigation for different diseases such as liver disease, leukaemia. Not yet, Congenital Heart Disease.

<mark>END OF THE WALKTHROUGH</mark>

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 condition?

b) Restrict the search based on targets for which the association with the disease was based on Somatic mutations. Does this list match 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 mutations (i.e. with a reference ID such as rsID) 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 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? Why should you expect this protein to have TM domains?

Exercise 2 – *GLP1R* and type II diabetes

BACKGROUND
GLP1R is a receptor for glucagon-like peptide 1, which is expressed in pancreatic beta cells. The activity of this receptor is mediated by G-proteins, which will lead to activation of the adenylyl cyclase pathway. This results in increased synthesis and release of insulin. For that reason, *GLP1R* has been under investigation as one of the possible drug targets to treat type II diabetes (Holst 2004).

SIGNIFICANCE
Type II diabetes can occur either when the body doesn't produce enough insulin to function properly, or the body's cells don't react to the hormone. This will result in higher levels of glucose in the blood and isn't used as fuel for energy. The number of people with diabetes has increased by three fold in a 34-year period. In 2012, an estimated 1.5 million deaths were directly caused by diabetes and another 2.2 million deaths were attributable to high blood glucose (source: WHO).

QUESTIONS
a) Are there other diseases within the broader therapeutic area of 'Metabolic disease' that associated with *GLP1R*? Can you name a few of them? Can you download this data as CSV (comma separated value)?

b) Which data types seem to confirm the association of *GLP1R* and type II diabetes?

c) How many unique drugs are currently mapped to this gene, which could potentially modulate target *GLP1R*? Are there any at phase IV (note: phase IV trails are carried out in drugs that are known to work

and for which there is a licence. In stage IV, side effects, safety of the drug, long-term benefits and risks will be further assessed).

d) Let's find out a bit more about the disease itself including its various synonyms, other drugs listed for the disease and its ontology. Where can you find this information using the Target Validation Platform?

ADDITIONAL EXERCISES:

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

Extra exercise 1 – Using the Target Validation Platform to find out if the modulation of a target by a drug poses any possible unsafety

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
Rofecoxib seems to be effective in the treatment of osteoarthritis, a degenerative disease of the joints. The target of this compound is *PTGS2.*

QUESTIONS
a) Does the Target Validation show an association between osteoarthritis and *PTGS2*?

b) Which pieces of evidence were used to support this association?

c) In which phase of the clinical trial is rofecoxib tested for the treatment of osteoarthritis? Are there other drugs under investigation as well, modulating the same target to treat osteoarthritis?

d) What is the level of baseline expression of the target (i.e. *PTGS2)* in normal heart?

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

Extra exercise 2 – 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
Franke et al (2006) have found three genes associated with congenital stationary night blindness:

ENSG00000163914
ENSG00000114349
ENSG00000133256

QUESTIONS
a) "How can I find out all diseases (besides Congenital stationary night blindness) 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.

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.

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

GENETIC VARIANTS and SOMATIC MUTATIONS
GWAS – The catalogue of Genome Wide Association Studies (GWAS) provides genetic variants (e.g. SNPs) that are associated with a disease.

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

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.

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

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.

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.

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.

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.

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.

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.