

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



**Hands-on Workshop
Course booklet**

**CRUK Cambridge Institute
16th December 2016**

**Denise Carvalho-Silva
Open Targets Outreach**

Notes

This workshop is based on the December 2016 release of the 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) Workshop materials (in pdf)

<https://github.com/deniseOme/training>

4) Our publication

nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

5) Details on the December release

<https://blog.opentargets.org/open-targets-platform-our-new-release-is-out-2/>

5) Feedback survey

<http://tinyurl.com/cam-161216>

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.....	4
INTRODUCTION TO OPEN TARGETS.....	5
OPEN TARGETS PLATFORM: WALKTHROUGH.....	8
HANDS-ON EXERCISES.....	25
EXTRA HANDS-ON EXERCISES	28
QUICK GUIDE TO DATABASES	32

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 fourth 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)
<http://nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055>

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.

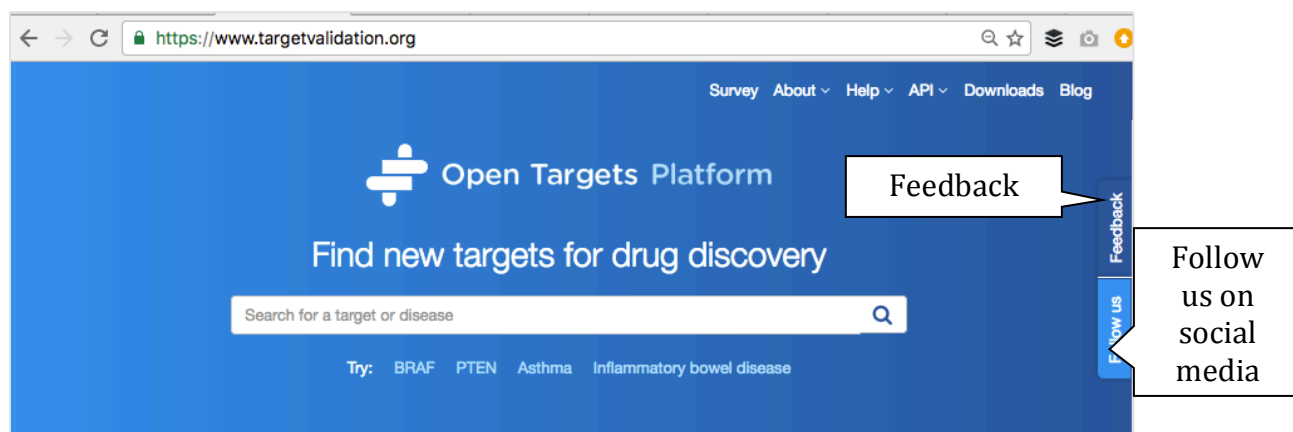
We will guide you through the website using congenital heart disease (CHD) as an example.

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".

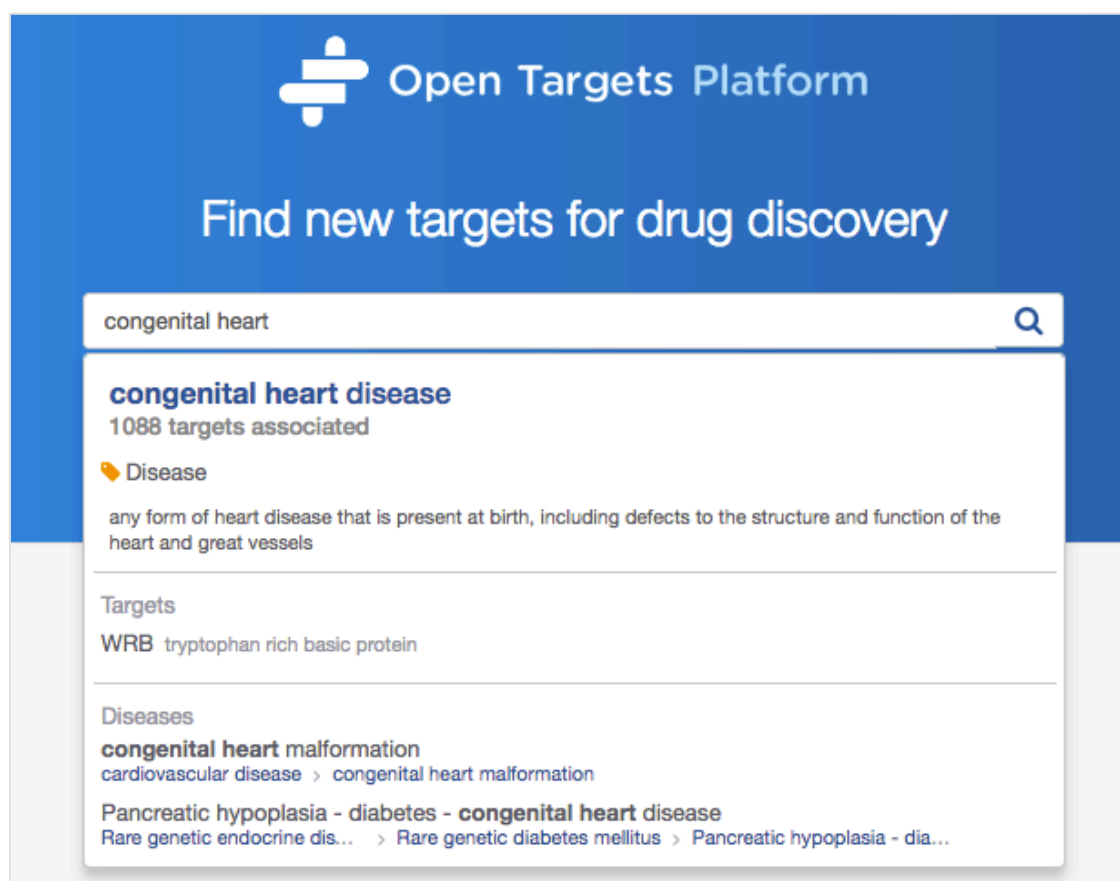
The following points will be addressed during the walkthrough:

- 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:

Search box

Total number of targets associated with CHD

Data types (Genetic Associations, Drugs, etc)

Filter the results

Pathway types (Immune system, metabolism, etc)

Target class (Enzyme, Transporter, etc)

Your target list

Download the table

1088 targets associated with congenital heart disease

Filter by

Clear all X Select all ✓

Genetic associations (24)

Somatic mutations (1)

Drugs (53)

Affected pathways (1)

RNA expression (0)

Text mining (838)

Animal models (135)

Pathway types

Clear all X Select all ✓

Signal Transduction (271)

Metabolism (209)

Immune System (205)

Developmental Biology (140)

Metabolism of proteins (137)

Hemostasis (120)

Disease (97)

Gene Expression (96)

Extracellular matrix organization (84)

Vesicle-mediated transport (81)

Transmembrane transport of s... (57)

Neuronal System (40)

Cell Cycle (40)

Muscle contraction (39)

Cellular responses to stress (37)

Organelle biogenesis and mai... (24)

Chromatin organization (18)

DNA Repair (17)

Cell-Cell communication (16)

Programmed Cell Death (16)

Target class

Clear all X Select all ✓

Enzyme (192)

Unclassified protein (71)

Membrane receptor (43)

Ion channel (39)

Secreted protein (32)

Transcription factor (19)

Other cytosolic protein (12)

Transporter (11)

Epigenetic regulator (10)

Adhesion (8)

Surface antigen (2)

Auxiliary transport protein (2)

Other nuclear protein (1)

Other membrane protein (1)

Target symbol

Association score

Genetic associations

Somatic mutations

Drugs

Affected pathways

RNA expression

Text mining

Animal models

Target name

Target class

Score

Previous 1 2 3 4 5 ... 22 Next

The table is sorted by default with the best hit on the top of the table. The best hit is for the target that contains the highest overall association score across the data types (Genetic associations, Somatic mutations, etc).

The sorting can also be done by alphabetical order of the list of targets or numerically according to 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 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):

Target symbol	Association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
GDF1									growth differentiation factor 1

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. Animal models.

The current release of the Open Targets Platform (December 2016) lists 1088 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

...

C) Target class

- Enzyme
- Membrane receptor

...

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

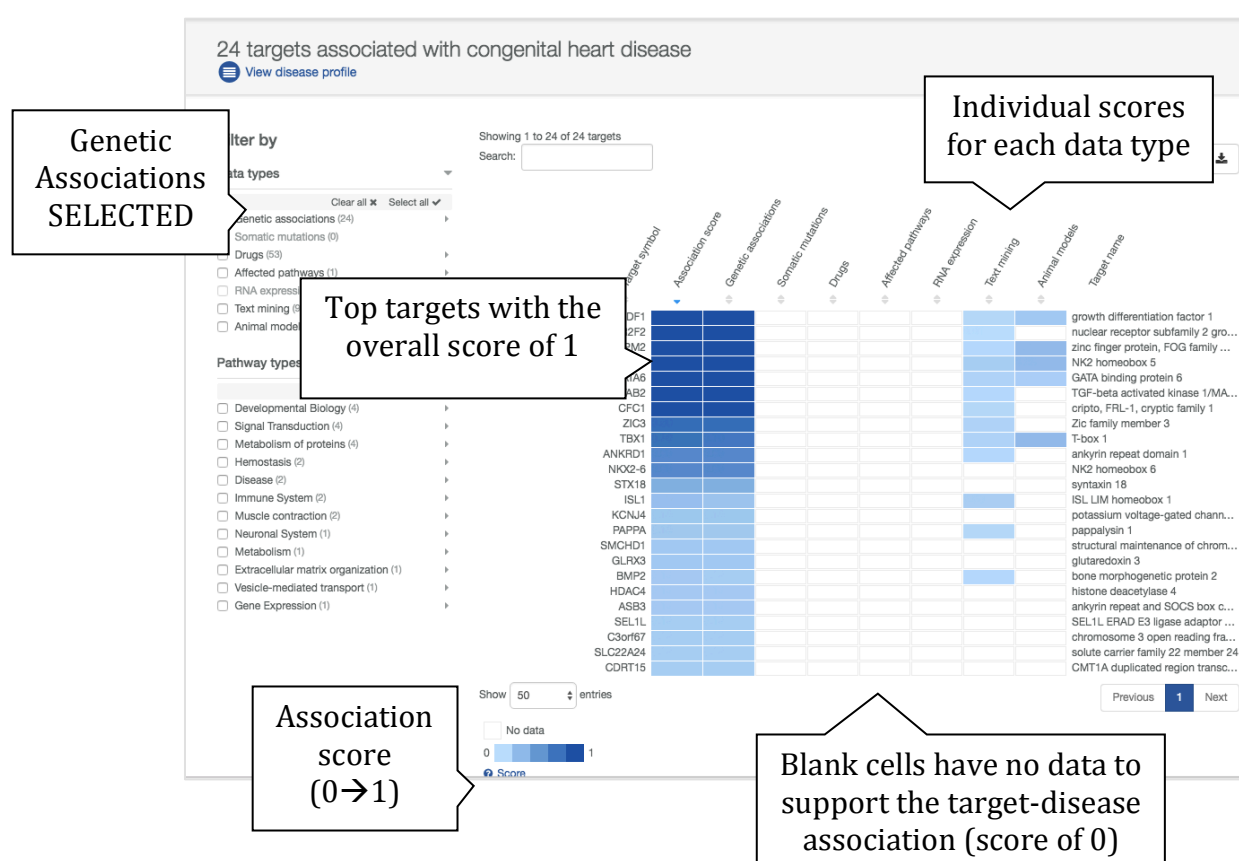
catalog and UniProt, both combined into Data types. Note that data from an individual source can contribute to different Data types, e.g. data from EVA is observed in two data types, Genetic associations and Somatic mutations.

Pathway types are defined by Reactome (<http://www.reactome.org/>).

The categories within Target class are defined by ChEMBL (<https://www.ebi.ac.uk/chembl/>).

Let's filter the data by 'Genetic associations' (Data type) to show the targets associated with CHD, which are based on genetic variants (such as SNPs from GWAS or UniProt data sources, for example) only.

The number of targets goes down to 24.



The individual data sources used to support the association of these targets to CHD within the Genetic association data type are:

GWAS catalog

UniProt
UniProt literature
European Variation Archive
Gene2Phenotype

Check our help page below to find out more about our data sources:

https://targetvalidation.org/data_sources

Alternatively, have a look at the Quick Guide to Databases at the end of this booklet.

Note: If you wish to suggest data or resources we could incorporate in our platform, please email them to 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, Text mining, Animal models) 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.

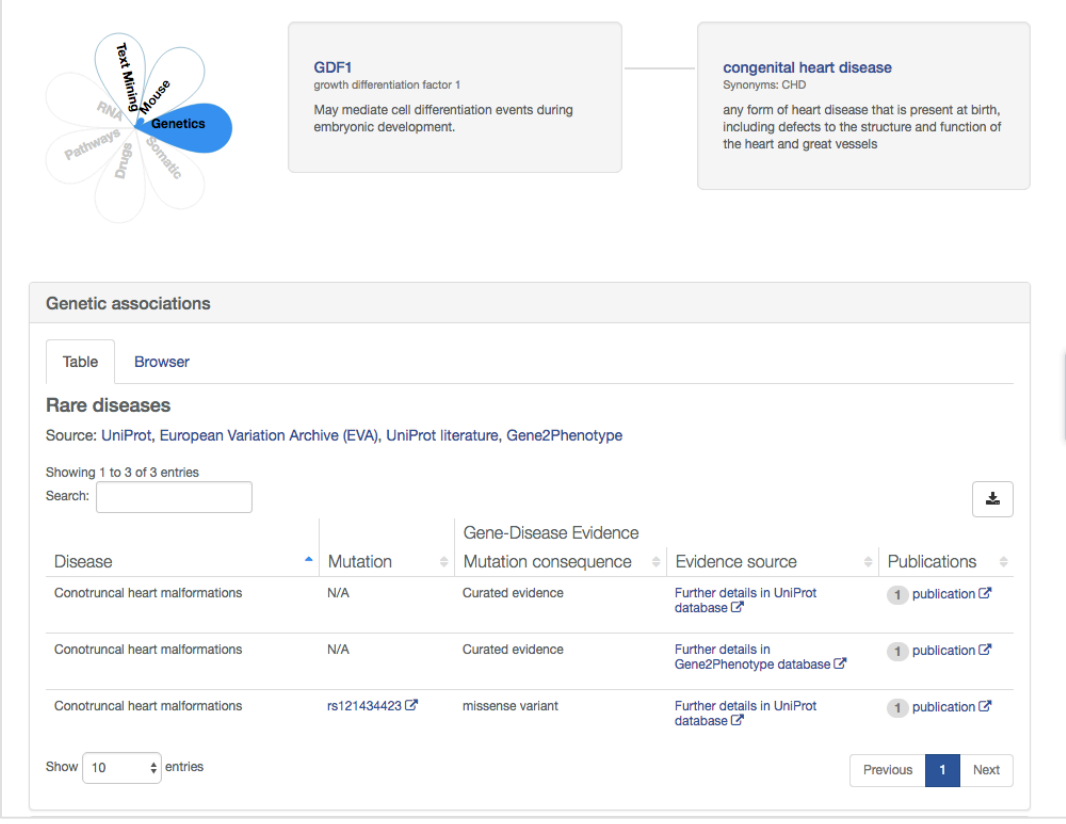
Alternatively, click on the cell in column 'Genetic associations' to land in the page below:

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 mutations, the Gene-Disease Evidence (e.g. curated), the mutation consequence (e.g. missense variant), the source of the data (e.g. UniProt, Gene2Phenotype) and relevant publications.

Check our Variants page for more details on the different mutation consequences: <http://targetvalidation.org/variants>



The interface displays genetic associations for the gene **GDF1** (growth differentiation factor 1) and the disease **congenital heart disease**. The GDF1 description states: "May mediate cell differentiation events during embryonic development." The congenital heart disease description states: "any form of heart disease that is present at birth, including defects to the structure and function of the heart and great vessels".

Genetic associations

Table | **Browser**

Rare diseases
Source: UniProt, European Variation Archive (EVA), UniProt literature, Gene2Phenotype

Showing 1 to 3 of 3 entries

Search:

Disease	Mutation	Gene-Disease Evidence	Evidence source	Publications
Conotruncal heart malformations	N/A	Curated evidence	Further details in UniProt database ↗	1 publication ↗
Conotruncal heart malformations	N/A	Curated evidence	Further details in Gene2Phenotype database ↗	1 publication ↗
Conotruncal heart malformations	rs121434423 ↗	missense variant	Further details in UniProt database ↗	1 publication ↗

Show entries

Previous **1** Next

Only one of these variants is known in public databases: its ID is rs121434423.

You can also explore Genetic associations in a Browser view: you can 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, both to the genomic regions and gene page.

Genetic associations

Table Browser

Human Chr: 19

human 19:18863034-18901606

Mutations in common diseases

Mutations in rare diseases

sequence

Genes / Transcripts

<CERS1

<GDF1-001

AC005197.2

<COPE

Click on the lollipop for more details on the variants

rs200024180

Ancestral allele	G
Allele string	G/A/C
Most severe consequence	missense_variant
MAF	0.000199681
Location	Jump to sequence
target	CERS1

Associations

Progressive myoclonic epilepsy	2 articles
Progressive myoclonic epilepsy	2 articles
Progressive myoclonic epilepsy	2 articles

Gene legend: ■ protein coding ■ antisense

Variants legend:

- Mutation in GDF1 associated with congenital heart disease
- Mutation associated with congenital heart disease in other genes
- Variant in GDF1
- Other Variant

Open region in Ensembl [↗](#) See GDF1 in 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 that support the association:

Text mining

Source: Europe PMC

Shown are the top 3 articles where **target** and **disease** are found in the same sentence.

Showing 1 to 3 of 3 entries

Search:

Disease Publication Year

congenital heart disease	<p>Association of GDF1 rs4808863 with fetal congenital heart defects: a case-control study.</p> <p>Zhang J et al. BMJ Open 5(12):e009352</p> <p>2015</p> <p>Abstract</p> <p>Congenital heart defects (CHDs) are the most common fetal defects and the most important cause of child mortality and morbidity. To investigate the association between growth/differentiation factor 1 (GDF1) polymorphisms and fetal CHDs, by evaluating the association of GDF1 rs4808863 with fetal CHDs. A case-control study. Beijing, China. We selected 124 fetuses with a CHD and a normal karyotype and normal array-based comparative genomic hybridisation analysis and compared them with 124 normal fetuses matched for gestational age and sex. Fetuses with a CHD, from 20 to 32 weeks of gestation were included. Fetuses with any chromosomal abnormalities, and fetuses from multiple pregnancies and those carried by pregnant women with chronic diseases, were excluded from this research. DNA extraction and genotyping were carried out for all cases to investigate the genotype distributions of GDF1 rs4808863. A significant difference was noted for the CT phenotype of GDF1 rs4808863 between the controls and the fetuses with CHDs using homozygote and heterozygote comparisons. The minor allele (T allele) of GDF1 rs4808863 was associated with an increased risk of CHD (p<0.05). A statistically significant difference between controls and fetuses with CHDs was noted in a comparison with the mutation genotype CT+TT and wild-type genotype CC (p<0.05) using dominant modal analysis. After stratification analysis, the CT phenotype, the minor allele (T allele) and the mutation genotype CT+TT of the rs4808863 polymorphism were associated with atrioventricular septal defect (AVSD), left ventricular outflow tract obstruction (LVOTO) and left-right laterality defects (p<0.05). Our results suggest that the GDF1 rs4808863 polymorphism contributes to an increased risk of fetal CHDs, especially the subtypes of AVSD, LVOTO and left-right laterality defects.</p>
--------------------------	---

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

Let's now explore the target in more detail. Still on the Evidence page, scroll back up and click on the GDF1 link next to the flower:

Evidence for GDF1 in congenital heart disease




GDF1
growth differentiation factor 1

g embryonic

Click on GDF1 for more details on this target

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:

Open Targets About ▾ Help ▾ API Downloads Blog

GDF1
growth differentiation factor 1 |  [View associated diseases](#)

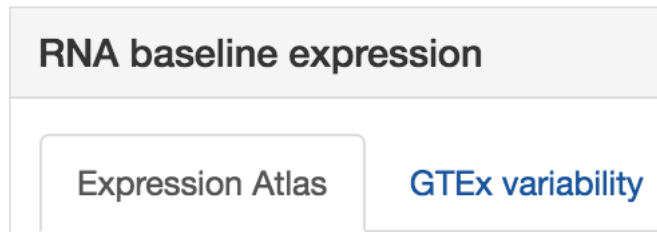
Click here to view all diseases associated with GDF1

May mediate cell differentiation events during embryonic development.

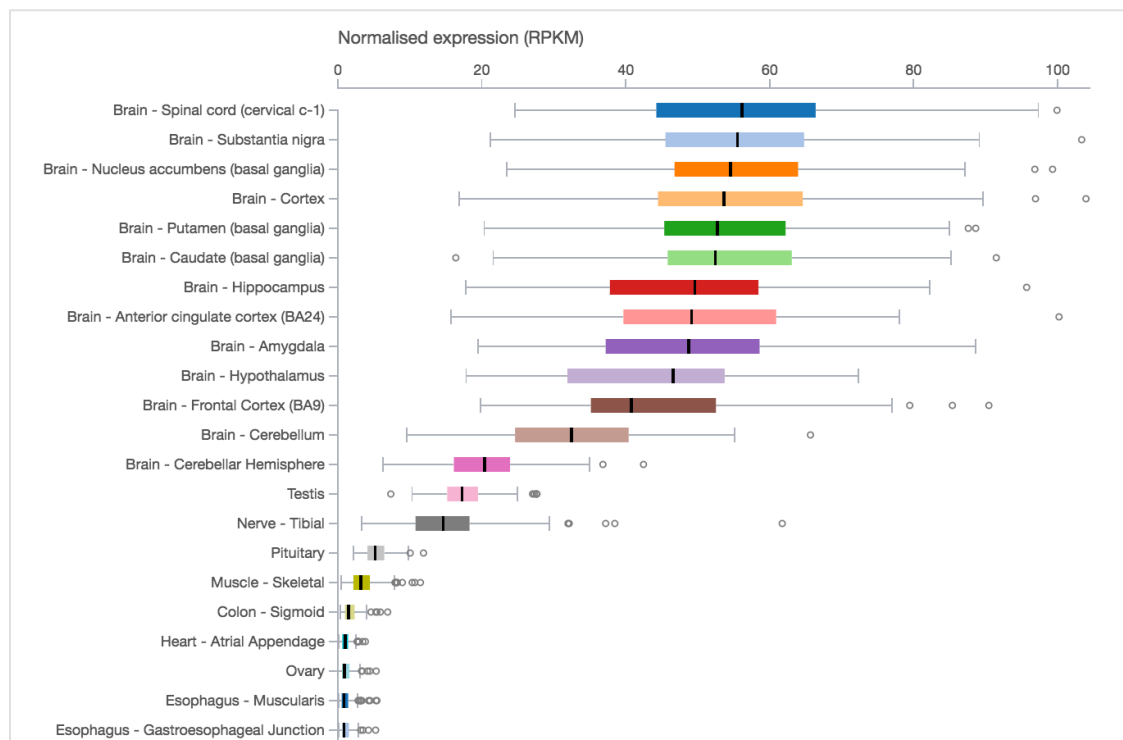
Synonyms: **GDF-1** Embryonic growth/differentiation factor 1

- Protein Information (from UniProt)
- Variants, isoforms and genomic context
- Protein baseline expression
- RNA baseline expression
- Gene Ontology
- Protein Structure
- Pathways
- Drugs

Expand the RNA baseline expression to find out in which tissues *GDF1* seems to be highly expressed. You have two tabs there, one with Expression Atlas data (including several projects), and one with GTEx data only:

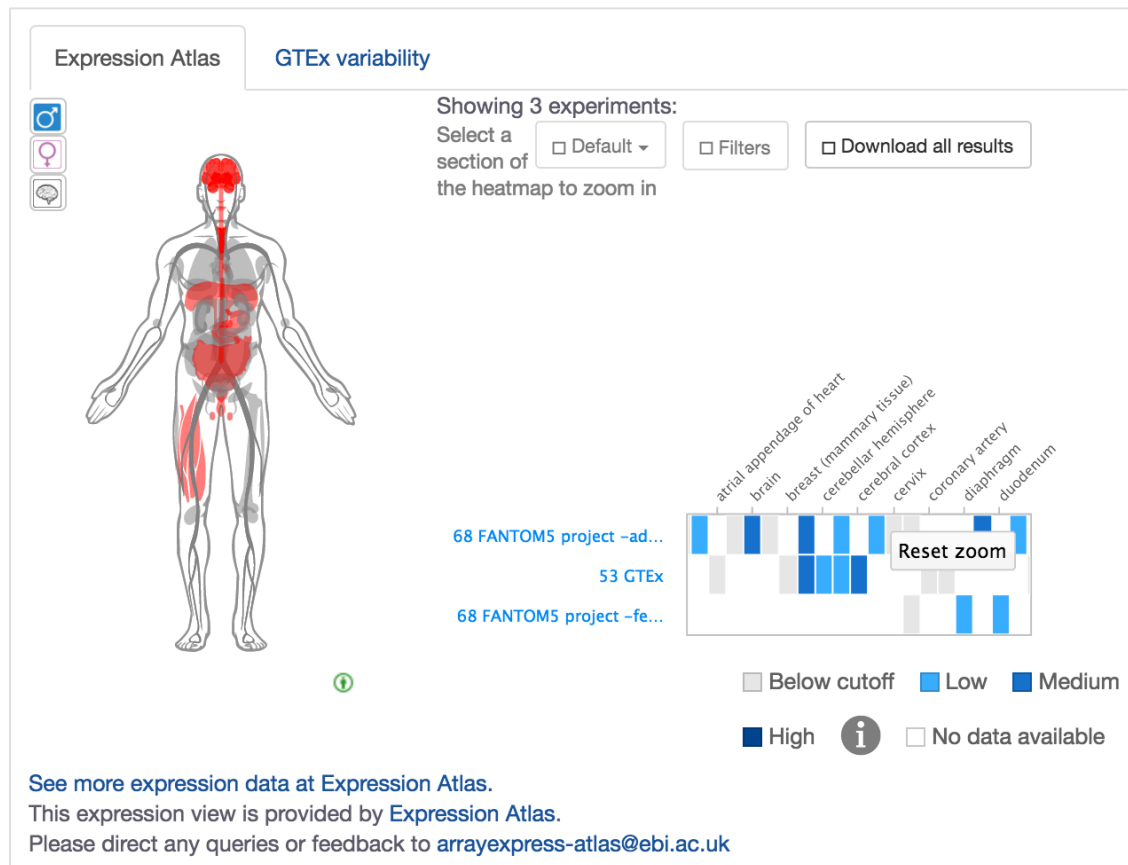


Let's have a look at the variability of expression data (in RPKM, Reads Per Kilobase Million, normalised for sequencing depth and gene length) from GTEx:

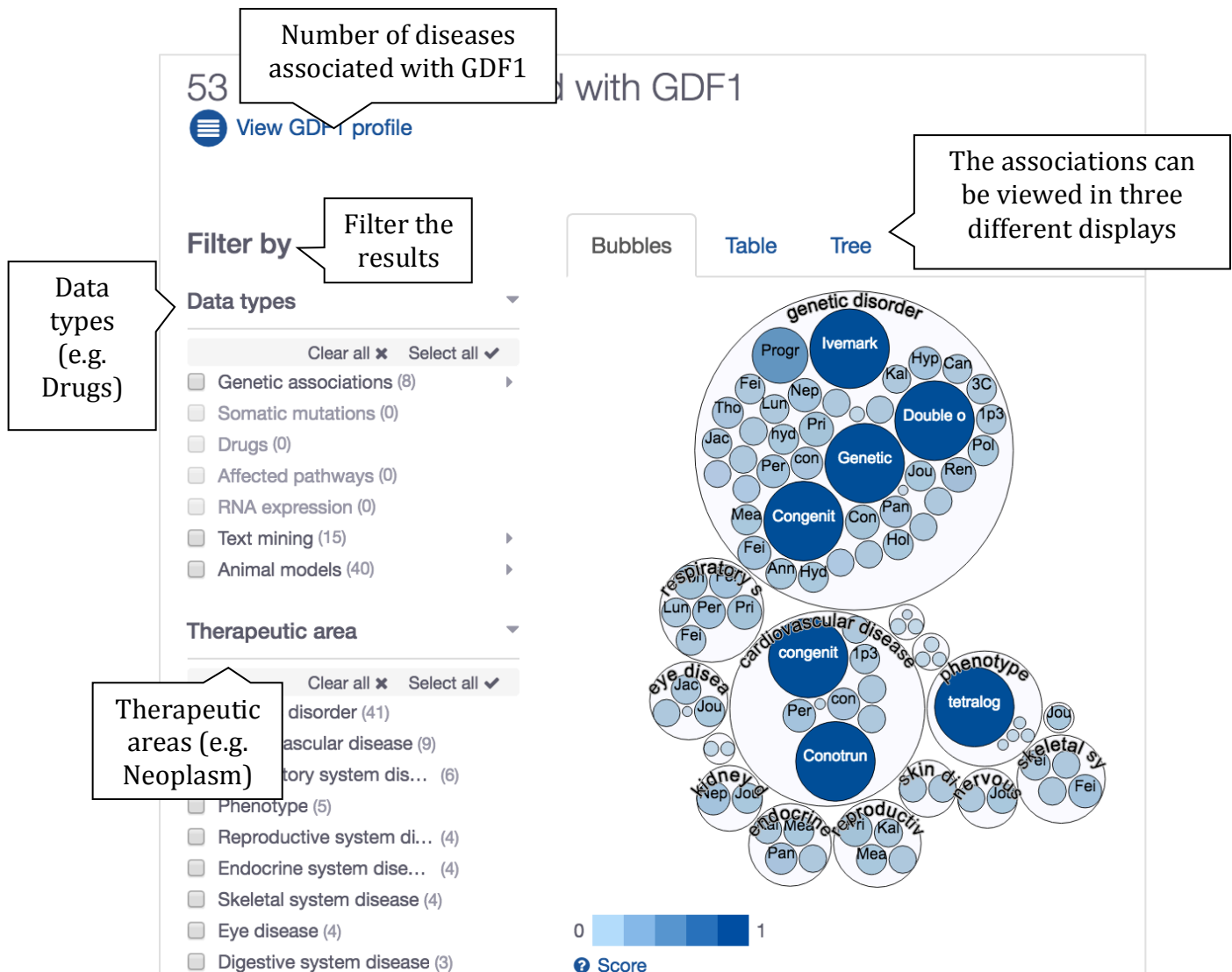


According to the GTEx plot, the spinal cord (brain) is the tissue with the highest expression levels for *GDF1*. The variability of the expression is shown and includes the median RPKM value.

You can also explore the Expression Atlas diagram with data from the Fantom5 project (adult and fetal tissues) and GTEx:



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



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

- Bubble view

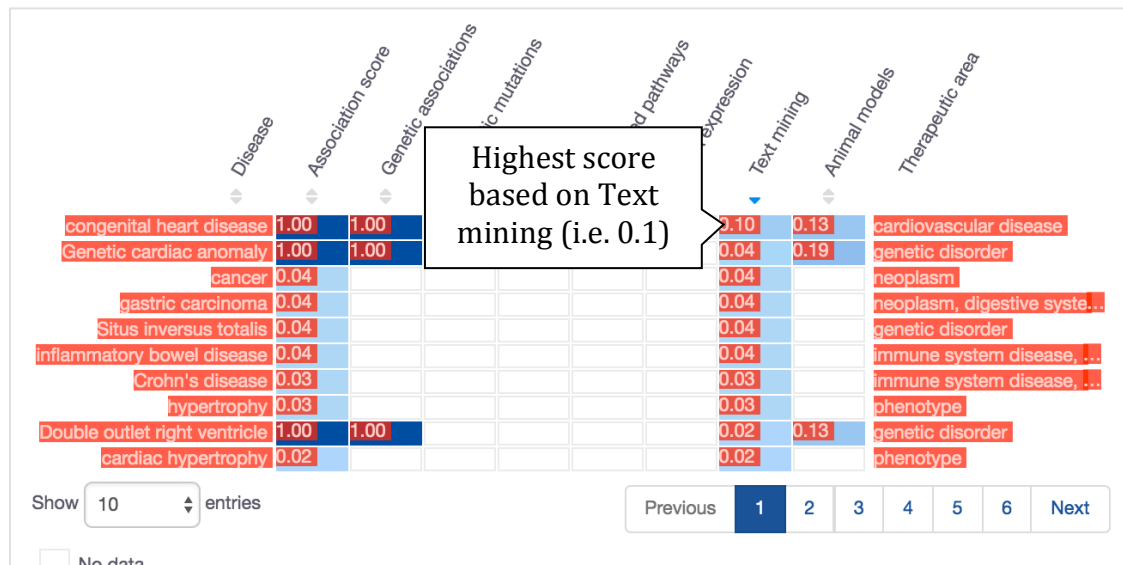
In this view, we group diseases into 'bubbles' based on the disease ontology. Large bubbles correspond to a therapeutic area and consist of smaller bubbles representing diseases within this area. A disease can belong to several therapeutic areas and therefore can appear within more than one large bubble. 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). You can show the 10 first entries and get the pagination for the remaining entries:

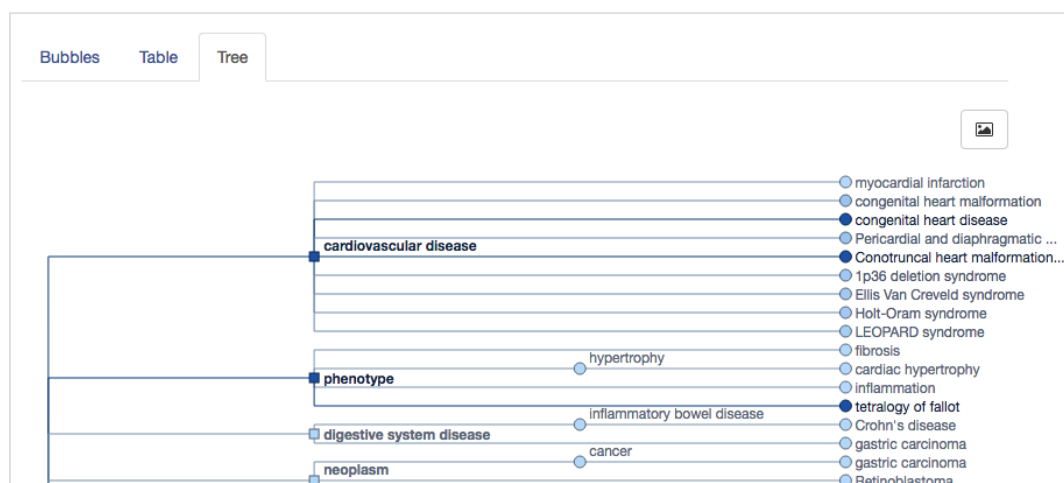


Tip: We colour code the cells in the table in different shade of blue as a visual way to convey the strength of the association (strongest association is coloured in dark blue). The score varies from 0 to 1. Hover over the cells in the table to view the numbers. Alternatively, you can select the cells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



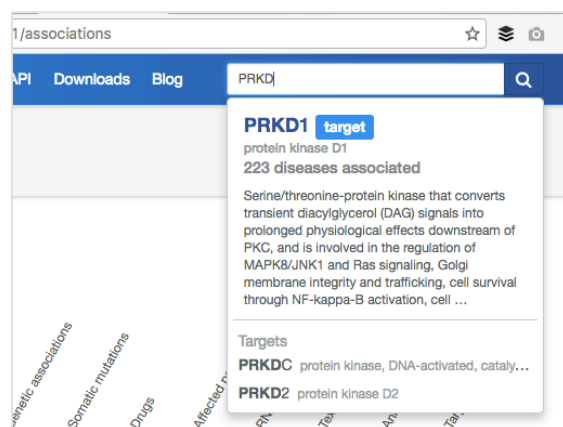
Regardless the view you choose to explore, you can filter the data to find out if there are other “Cardiovascular diseases” associated with this target. There are nine of them (including coronary heart disease, CHD): Conotruncal heart malformations (with the highest score of 1) plus a few others with 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 Platform by starting off with a disease (i.e. congenital heart disease, or CHD).

You can also search for genes, drugs and phenotypes.

Let's have a look at the *FOXP2* gene, one of the targets described in the paper "Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing" 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 here:

www.targetvalidation.org

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

← → ↻ <https://www.targetvalidation.org/target/ENSG00000184304>

 Open Targets About ▾ Help ▾ API

223 diseases associated with PRKD1

 [View PRKD1 profile](#)

Click on 'View PRKD1 profile' for more information on this gene


Filter by Bubbles Table Tree

Data types ▾

Clear all ✕ Select all ✓

- ☐ Genetic associations (5) ▸
- ☐ Somatic mutations (0)
- ☐ Drugs (16) ▸

Let's now click on the 'View PTPN11 profile'. Then on 'Drugs':

PRKD1
protein kinase D1 |  [View associated diseases](#)

Serine/threonine-protein kinase that converts transient diacylglycerol into a permanent signal by phosphorylating MAPK8/JNK1 and Ras signaling, G-protein coupled receptor (GPCR) signaling, cell migration, cell differentiation by mediating HDAC7 nuclear export, VEGFA-induced angiogenesis, genotoxic-induced apoptosis, epidermal growth factor receptor (EGFR) on dual threonine residues, MAPK8/JNK1 activation and subsequent JUN phosphorylation. Farnesyl transferase (FTase) and increased competition with RAF1 for binding to GTPase-activating protein (GAP) heterotetramer. [show more]

Synonyms: PKCM PKD PKC-mu PKD1 PRKCM 2.7.11.13
Protein kinase C mu type

- Protein Information (from UniProt)
- Variants, isoforms and genomic context
- Protein baseline expression
- RNA baseline expression
- Gene Ontology
- Protein Structure
- Pathways
- Drugs**

Expand this tab to view the Drugs mapped to target *PRKD1*

You will find out which drugs are currently in clinical trials and could be modulating this gene:

Drugs								
Source: ChEMBL								
Found 4 unique drugs: GSK-690693 MIDOSTAURIN SOTRASTAUIN UCN-01								
Showing 1 to 10 of 15 entries								
Search: <input type="text"/>								
Drug Information							Gene-Drug Evidence	
Disease	Drug	Phase	Status	Type	Mechanism of action	Activity	Target class	Evidence source
posterior uveitis	SOTRASTAUIN	Phase II	Completed	Small molecule	Protein kinase C (PKC) inhibitor 1 publication	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information
psoriasis	SOTRASTAUIN	Phase II	Completed	Small molecule	Protein kinase C (PKC) inhibitor 1 publication	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information
acute myeloid leukemia	MIDOSTAURIN	Phase II	Recruiting	Small molecule	Protein kinase C (PKC) inhibitor 1 publication	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information
melanoma	UCN-01	Phase II	Terminated	Small molecule	Protein kinase C (PKC) inhibitor 1 publication IUPHAR	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information
ulcerative colitis	SOTRASTAUIN	Phase II	Terminated	Small molecule	Protein kinase C (PKC) inhibitor 1 publication	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information
acute myeloid leukemia	MIDOSTAURIN	Phase II	Not yet recruiting	Small molecule	Protein kinase C (PKC) inhibitor 1 publication	antagonist	AGC protein kinase PKC alpha subfamily	Curated from Clinical Trials Information

The source for this data is from ChEMBL. There are four unique drugs mapping to this target, i.e. SOTRASTAUIN, MIDOSTAURIN and GSK-690693, and UCN-01.

The same drug can be at different phases of the clinical trial process for different types of diseases (including cancer).

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 Somatic mutations only. Are there any difference between this list and 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 in 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) Can you list some of the pathways this target is associated with?

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 'Hematological system' are associated with this target? How many of these are based on 'Drugs' only?

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? Are there any drugs in clinical phase IV status completed for this disease? What are the differences you see in the results after the filter is applied?

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 (the 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 EGFR gene, a receptor tyrosine kinase

BACKGROUND

EGFR is a cell surface protein that binds to epidermal growth factor. Binding of the protein to a ligand induces receptor dimerization and tyrosine autophosphorylation and leads to cell proliferation. Mutations in this gene are associated with lung cancer.

QUESTIONS

- a) How long is the protein encoded by this gene/target?*
- b) Which amino acids correspond to the kinase domain? Are there other domains and/or sites mapped to this protein?*
- c) Which tissue has the highest RNA baseline expression from*
 - c.1 the GTEx project*
 - c.2 the ENCODE project in Expression Atlas*
- d) Can you list a few examples of molecular functions the EGFR 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. Can you use the Open Targets Platform to find out where there is a mouse orthologue for this human gene?*
- f) Which drugs are currently in clinical trial IV, status completed, targeting this gene in non-small cell lung carcinoma?*
- g) Are these drugs targeting other genes than EGFR? (tip: you will be searching for these drugs in the Open Targets Platform)*

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 kinase class of target has the strongest association score for non-small cell lung carcinoma based on evidence for somatic mutations only?*
 - b) Select the first gene in the table. Can you find out whether the 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?*
 - d) Deselect all filters previously selected. You should have 5774 targets associated with non-small cell lung carcinoma now. Can you filter this list with the text file exercise4.txt? How many genes in the exercise4 file match the original list?*
-

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. CDK6 has been shown to phosphorylate and thus regulate the activity of tumor suppressor protein Rb. Expression of this gene is up-regulated in some types of cancer.

Abemaciclib is under investigation in patients with breast carcinoma among other types of cancer.

QUESTIONS

a) Which data supports the association between CDK6 and breast carcinoma?

b) Are there other drugs in addition to abemaciclib used in clinical trials modulating the same target for breast carcinoma? Can you get to the original data?

c) Are there studies showing a decreased level of RNA expression of this gene in breast carcinoma?

d) What is the level of RNA baseline expression of the target (i.e. CDK6) in heart according to the '19 NIH Epigenomics Roadmap' project?

e) Is this target associated with cardiovascular diseases with a strong confidence (i.e. score of 0.80 or above)?

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 using our REST API (or Python, 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 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/>