

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



**Hands-on Workshop
Coursebook**

**Yale School of Medicine
9th February 2017**

**Denise Carvalho-Silva
Open Targets Outreach**

Notes

This workshop is based on the December 2016 release of our 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

www.bit.ly/OpenTargets

5) Details on the latest Platform release

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

6) Feedback survey

<http://tinyurl.com/yale-090217>

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@targetvalidation.org

TABLE OF CONTENTS

OVERVIEW.....	4
INTRODUCTION TO OPEN TARGETS.....	5
OPEN TARGETS PLATFORM: LIVE DEMOS.....	8
Demo 1.....	9
Demo 2.....	16
Demo 3.....	28
HANDS-ON EXERCISES.....	24
Exercise 1.....	24
Exercise 2.....	25
Exercise 3.....	26
EXTRA HANDS-ON EXERCISES	33
Exercise 4.....	33
Exercise 5.....	33
Exercise 6.....	35
QUICK GUIDE TO DATABASES	36

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 identification of targets relevant to key therapeutic areas namely:

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

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

More details can be found in our [Projects](#) page.

Retrieving data from Open Targets with our Platform

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

In our release (December 2016), the Platform provides information on 31,071 targets; 8,659 diseases; 4.9 million evidence; and 2.5 million target-disease associations.




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

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

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

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

Help documentation and support

-  [Data sources](#) in the Open Targets Platform
-  View our [FAQs](#)
-  [Email us](#)

Connect with us

- ❖ [Open Targets Blog](#)
- ❖ Follow us on [Twitter](#)
- ❖ Check our page on [Facebook](#) and [LinkedIn](#)

Further reading

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

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

OPEN TARGETS PLATFORM: LIVE DEMOS

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.

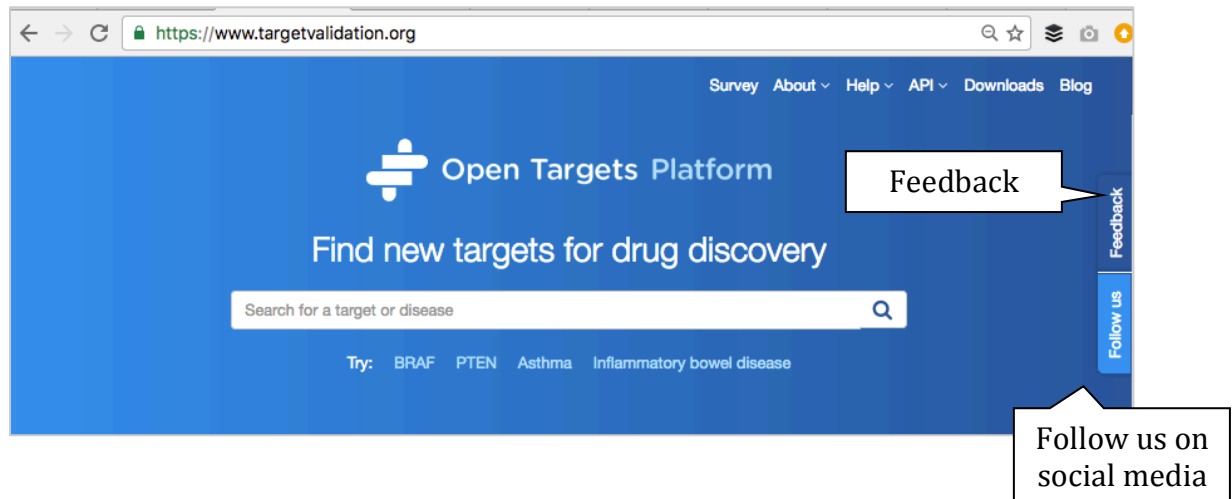
We will guide you through the website using renal cell carcinoma, (papillary) as an example and exploring the *MET* gene based on 'Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes' by Franke et al. AJHG 2006).

The following points will be addressed during the walkthrough:

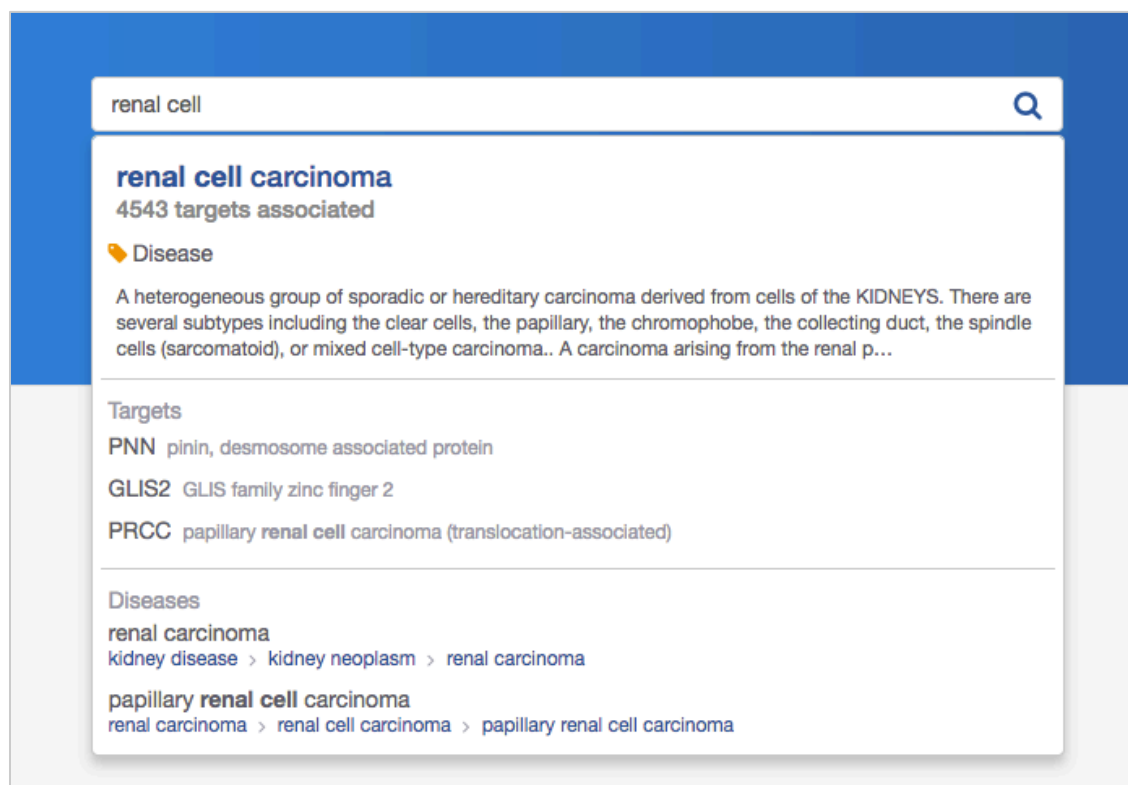
- Targets associated with CHD
- Filter down the number of targets based on specific evidence
- Data sources used to support the target-CHD association
- Looking for other diseases associated than CHD with a target
- Visualise a target in a browser like view
- Find out how strong is the association between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

Demo 1

Go to www.targetvalidation.org and search for one of the diseases described in the paper (renal cell carcinoma).



Select the first (best) hit:



You will see a page like this:

Total number of targets associated with renal cell carcinoma

4543 targets associated with renal cell carcinoma

Data types (Genetic Associations, Drugs, etc)

Filter the results

Filter by

1 to 50 of 4,543 targets

Target symbol	Association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
MET									MET proto-oncogene, rec...
VHL									von Hippel-Lindau tumor ...
PBRM1									polybromo 1
PDGFRA									platelet derived growth fa...
RET									ret proto-oncogene
PDGFRB									platelet derived growth fa...
FGFR3									fibroblast growth factor re...
BRAF									B-Raf proto-oncogene, s...
FLT4									fms related tyrosine kinas...

The current release of the Open Targets Platform (December 2016) lists 4543 targets associated with renal cell carcinoma.

The data types supporting these results are based on Genetic association, Somatic mutations, Drugs, RNA expression, Text mining, and Animal models.

Note: We will go through these data types (and the data sources that make the data types) later in this tutorial.

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

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

A) Data types

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

B) Pathway types

Signal Transduction

Metabolism

...

C) Target class

Enzyme

Membrane receptor

...

D) Your target list

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

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


We collect data from various sources and combine them into categories called Data types. Example of data sources are GWAS catalog and UniProt, both combined into Data types. Note that data from an individual source can contribute to different Data types, e.g. data from EVA is observed in two data types, Genetic associations and Somatic mutations.

'Pathway types' are defined by Reactome:


(<http://www.reactome.org/>),

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

Let's now filter the data to focus on 'Genetic associations' (Data type) only. The number of targets goes down to 15:

Open Targets Platform Survey About ▾ Help ▾ API ▾ Downloads Blog 

15 targets associated with renal cell carcinoma

 [View disease profile](#)

Filter by Showing 1 to 15 of 15 targets Search:

Data types Clear all ✕ Select all ✓

- ☒ Genetic associations (15)
 - ☐ GWAS catalog (8)
 - ☐ UniProt literature (5)
 - ☐ European Variation ... (4)
 - ☐ UniProt (3)
- ☐ Somatic mutations (526)
- ☐ Drugs (94)
- ☐ Affected pathways (0)
- ☐ RNA expression (3k)
- ☐ Text mining (2k)
- ☐ Animal models (1)

Pathway types

Top targets with the overall score of 1

Target symbol	Association score	Genetic associations	Somatic mutations	Drugs
PBRM1	1	1	0	0
HNF1A	1	1	0	0
SETD2	1	1	0	0
FLCN	1	1	0	0
EPAS1	1	1	0	0
MYC	1	1	0	0
HNF1B	1	1	0	0

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

These are targets associated with renal cell carcinoma based on genetic variants only. Genetic variants are SNPs from the GWAS catalog, UniProt and EVA (European Variation Archive) sources only.

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

Genetic association = Somatic mutations = drugs = pathways > RNA expression > Animal models > Text mining (lowest weight of all data types).

You can sort the table by alphabetical order of the list of targets, or by the association score (either overall or per data type e.g. Genetic associations, Drugs, Text mining, etc). The association score varies

from 0 to 1, the closer to 1 the stronger the association. This score is computed for each piece of evidence that is used to support the association 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
MET									MET proto-oncogene, rec...
VHL									
PBRM1									


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

Click on the arrows to sort the results by score values of individual data types e.g. Animal models.

Note: More details on the scoring will be given towards the end of this tutorial.

To see which drugs are used in the treatment of this disease, we need to click on the 'View disease profile':

15 targets associated with renal cell carcinoma

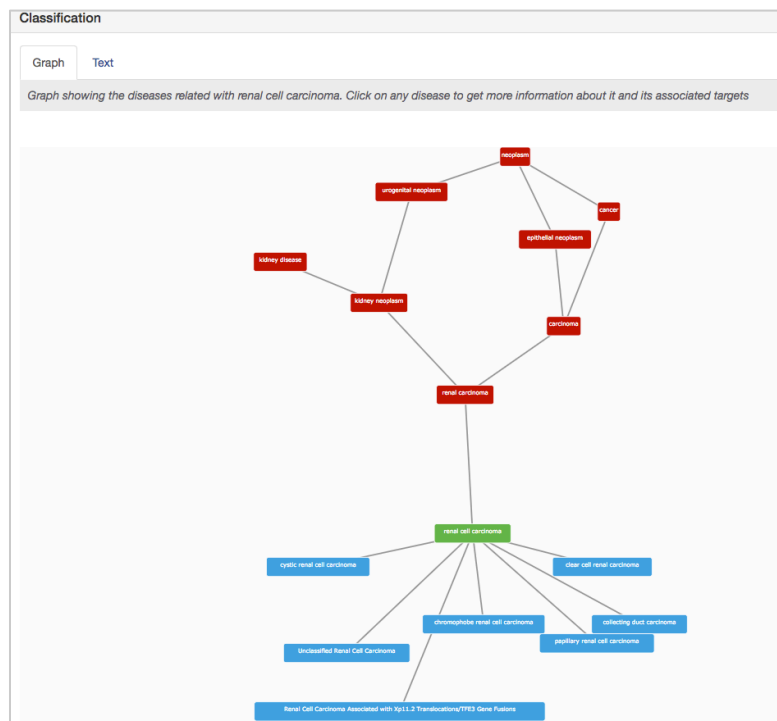
 [View disease profile](#)

Let's now expand the tab 'Drugs' to get a list of all drugs (n = 35 unique drugs in the December release) in different phases of clinical trials used in patients suffering from renal cell carcinoma. You can

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

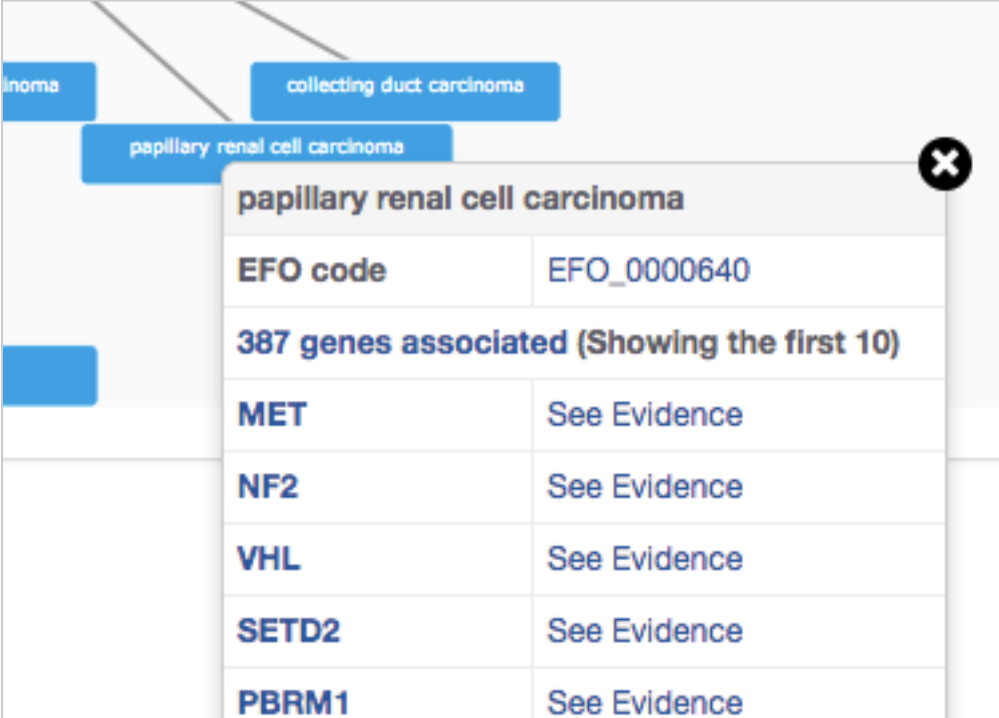
Drugs								
Source: ChEMBL								
Found 35 unique drugs: AFLIBERCEPT ALDESLEUKIN APITOLISIB AXITINIB AZD-2014 Anlotinib BEVACIZUMAB CABOZANTINIB CEDIRANIB DOVITINIB ERLOTINIB EVEROLIMUS Famitinib GIRENTUXIMAB INTERFERON ALFA-2A IXABEPILONE LINIFANIB MK-2206 NINTEDANIB NIVOLUMAB PANOBINOSTAT PAZOPANIB ROMIDEPSON SEMAXANIB SORAFENIB SUNTINIB Savolitinib TANDUTINIB TEMSIROLIMUS Tivantinib TIVOZANIB VANDETANIB VOLOCIXIMAB VORINOSTAT ZOLEDRONIC ACID								
Showing 1 to 10 of 1,000 entries								
Search: <input type="text"/>								
Drug Information							Gene-Drug Evidence	
Disease	Drug	Phase	Status	Type	Mechanism of action	Activity	Target class	Evidence source
renal cell carcinoma	EVEROLIMUS ↗	Phase IV	Recruiting	Small molecule	FK506-binding protein 1A inhibitor DailyMed ↗	antagonist	Isomerase	Curated from Clinical Trials Information ↗
renal cell carcinoma	EVEROLIMUS ↗	Phase IV	Completed	Small molecule	FK506-binding protein 1A inhibitor DailyMed ↗	antagonist	Isomerase	Curated from Clinical Trials Information ↗
renal cell carcinoma	EVEROLIMUS ↗	Phase IV	Recruiting	Small molecule	FK506-binding protein 1A inhibitor DailyMed ↗	antagonist	Isomerase	Curated from Clinical Trials Information ↗
clear cell renal carcinoma	AXITINIB ↗	Phase IV	Recruiting	Small molecule	Vascular endothelial growth factor receptor inhibitor DailyMed ↗	antagonist	Tyrosine protein kinase VEGFR family	Curated from Clinical Trials Information ↗

Scroll down to view the disease ontology (disease relationship) in the 'Classification' tab:



Renal cell carcinoma is the disease of interest in this tutorial, and is represented in green. Red nodes correspond to parental terms in relation to renal cell carcinoma, whereas its children terms (e.g.

papillary renal cell carcinoma) are shown in blue. Click on any of disease names to get the targets associated with them:



papillary renal cell carcinoma

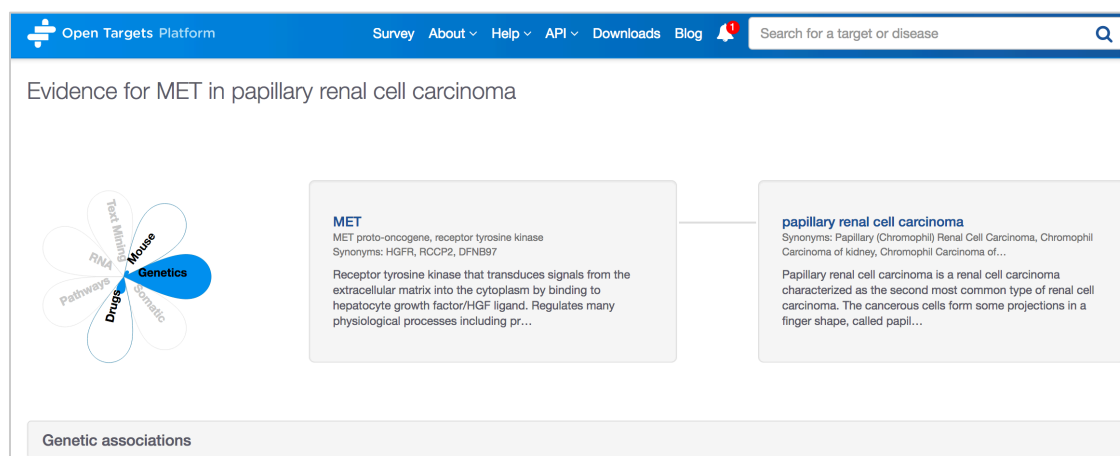
EFO code	EFO_0000640
387 genes associated (Showing the first 10)	
MET	See Evidence
NF2	See Evidence
VHL	See Evidence
SETD2	See Evidence
PBRM1	See Evidence

By using the EFO parent-child (subclass of) relationships, we derive new associations that may not have direct evidence. For instance, IBD is an autoimmune disease and the direct evidence of targets associated to IBD are propagated to the higher autoimmune level to allow users to find common targets across groups of related diseases (e.g. Ulcerative Colitis, Crohn's disease and IBD). In EFO, 'asthma' is a 'respiratory system disease' and 'childhood onset asthma' is a subclass of 'asthma'. Both evidence from 'asthma' and 'childhood onset asthma' are propagated to 'respiratory system disease'

Demo 2

Let's now explore the evidence used to associate *MET* and papillary renal cell carcinoma.

Click on 'See Evidence' link at the end of part I of the walkthrough to will land on a page like this:



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

Genetic association
Drugs
Animal models

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

support@targetvalidation.org.

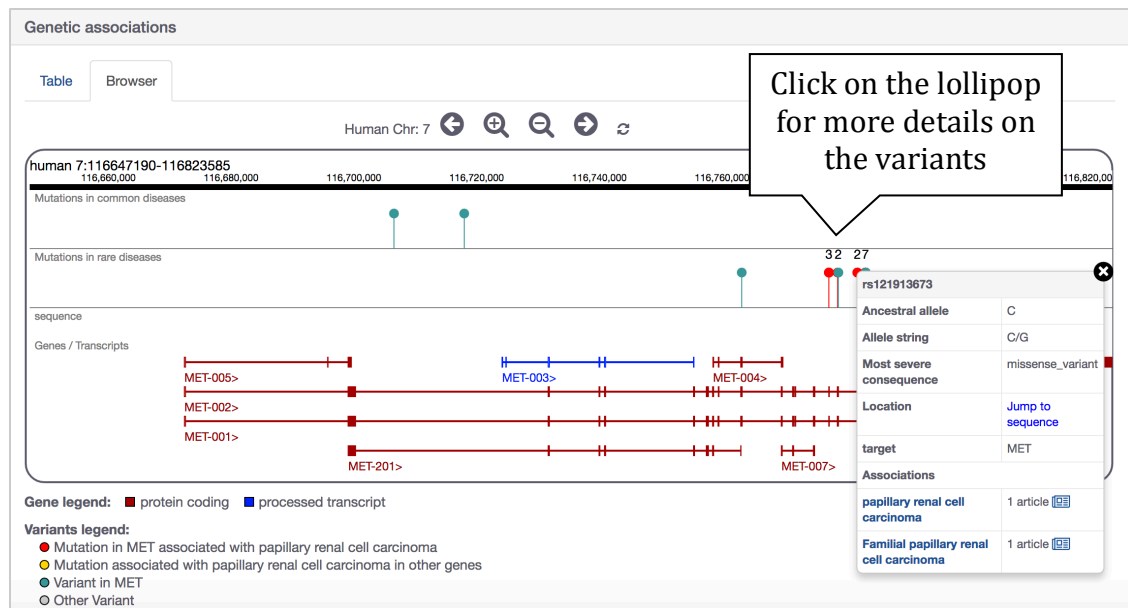
The genetic evidence that supports the *MET*-papillary renal cell carcinoma comes from UniProt. Most variants are known in public databases such as dbSNP (note the rsID):

Genetic associations				
<div>Table Browser</div>				
Rare diseases				
Source: UniProt, European Variation Archive (EVA), UniProt literature, Gene2Phenotype				
Showing 1 to 10 of 12 entries				
Search: <input type="text"/>				
Disease	Mutation	Gene-Disease Evidence	Evidence source	Publications
papillary renal cell carcinoma	N/A	Curated evidence	Further details in UniProt database	5 publications
papillary renal cell carcinoma	rs121913673	missense variant	Further details in UniProt database	1 publication
papillary renal cell carcinoma	rs121913670	missense variant	Further details in UniProt database	1 publication
papillary renal cell carcinoma	rs121913244	missense variant	Further details in UniProt database	1 publication

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

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

We also provide links to Ensembl.



To find out if there are drugs currently in clinical trials in patients with papillary renal cell carcinoma, let’s expand the ‘Drugs’ tab. There is just one drug (i.e. TIVANTINIB) on phase II of clinical trials.

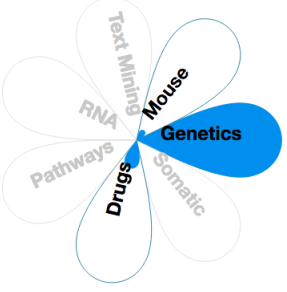
Let’s now find out the mouse model with phenotypes that mimic the human disease. Still on the same page, scroll down to view the ‘Animal models’ tab. Click on it to see there is one mouse model mutated at

this gene, which has the following phenotypes: increased sarcoma incidence, glomerulonephritis and hydronephrosis:

Animal models				
Source: Phenodigm				
Showing 1 to 1 of 1 entries				
Search: <input type="text"/>				
Disease	Phenotype - Phenotype Evidence		Model - Disease Evidence	Source
	Human	Mouse	Mouse model	
papillary renal cell carcinoma	<ul style="list-style-type: none"> Papillary renal cell carcinoma 	<ul style="list-style-type: none"> increased sarcoma incidence glomerulonephritis hydronephrosis 	Met^{tm40ny}/Met⁺ Involves: 129 * C57BL/6J	Phenodigm
Show <input type="text" value="10"/> entries Previous 1 Next				

We can now scroll back to the top of the page and click on the ‘MET’ link (next to the flower) to explore this target in more details, such as to find out its RNA expression levels across several tissues:

Evidence for MET in papillary renal cell carcinoma



MET

MET proto-oncogene, receptor tyrosine kinase
 Synonyms: HGFR, RCCP2, DFNB97

Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to hepatocyte growth factor/HGF ligand. Regulates many physiological processes including pr...

You will land on a page like this:

[Open Targets Platform](#)
[Survey](#)
[About](#)
[Help](#)
[API](#)
[Downloads](#)
[Blog](#)

MET

MET proto-oncogene, receptor tyrosine kinase | [View associated diseases](#)

Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to hepatocyte growth factor (HGF) and its related peptides. Ligand binding at the cell surface induces autophosphorylation and provides docking sites for downstream signaling molecules. Following activation by ligand, interacts with the PI3-kinase/AKT pathway. Recruitment of these downstream effectors by MET leads to the activation of several signaling cascades including RAS-ERK activation is associated with the morphogenetic effects while PI3K/AKT coordinates prosurvival effects. During gastrulation, development and migration of muscles and neuronal precursors, angiogenesis and kidney formation. In adult life, MET is involved in the regulation of cell growth, differentiation and survival.

Synonyms: [HGFR](#) [RCCP2](#) [DFNB97](#) [Scatter factor receptor](#) [2.7.10.1](#) [HGF receptor](#) [Hepatocyte growth factor receptor](#) [SF receptor](#) [hepatocyte growth factor receptor](#) [Proto-oncogene c-Met](#)

Protein Information (from UniProt)

Variants, isoforms and genomic context

Protein baseline expression

RNA baseline expression

Gene Ontology

Protein Structure

Pathways

Drugs

Gene tree

Bibliography

Expand the ‘RNA baseline expression’ to find out in which tissues *MET* is highly expressed.

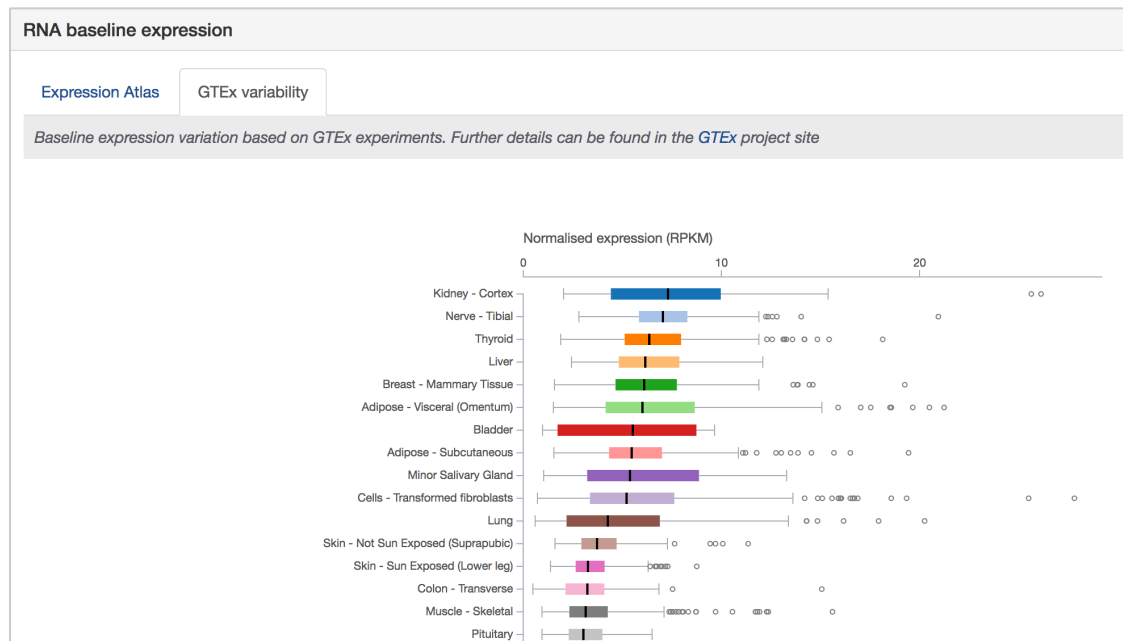
You will find two tabs in there: one with Expression Atlas data (including several projects), and one with GTEx data only:

RNA baseline expression

Expression Atlas

GTEx variability

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 the GTEx project:



According to the GTEx plot available in the Open Targets Platform, kidney (cortex) is the tissue with the highest expression level for the *MET* gene. The variability of the mRNA expression is shown and includes the median RPKM value.

Let's now click on 'View associated diseases' to find out all diseases associated with the *MET* (apart from papillary renal cell carcinoma):

Open Targets Platform Survey About ▾ Help ▾

MET
MET proto-oncogene, receptor tyrosine kinase

View associated diseases

You will land on a page like this:

Number of diseases associated with MET

169 diseases associated with MET

View MET profile

Filter by

Filter the results

Data types

Clear all ✕ Select all ✓

☐ Genetic associations (17)
☐ Somatic mutations (18)
☐ Drugs (43)
☐ Affected pathways (2)
☐ RNA expression (61)
☐ Text mining (114)
☐ Animal models (41)

Therapeutic area

Clear all ✕ Select all ✓

☐ Neoplasm (107)
☐ Genetic disorder (28)
☐ Other (26)
☐ Nervous system disease (24)
☐ Digestive system disease (19)
☐ Respiratory system disease (15)
☐ Endocrine system disease (15)

Bubbles

Table

Tree

The associations can be viewed in three different displays

Data types (e.g. Drugs)

Therapeutic areas (e.g. Neoplasm)

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

- Bubble view

In this view, we group diseases into 'bubbles' based on the disease ontology. Large bubbles correspond to a therapeutic area and consist of smaller bubbles representing diseases within this area. A disease can belong to several therapeutic areas and therefore can appear within more than one large bubble. The strength of the association between the target and a disease is represented by the size of the bubble and the shade of its blue colour; the larger the bubble and the darker the blue, the stronger the association.

- Table view

In this view, we list all diseases associated with a target, ordered by the association score, which is colour coded. When there is no evidence to support the association, the cells in this table are coloured

21

in white (score of zero). You can show the 10 first entries and get the pagination for the remaining entries.

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

BubblesTableTree

Showing 1 to 50 of 169 entries

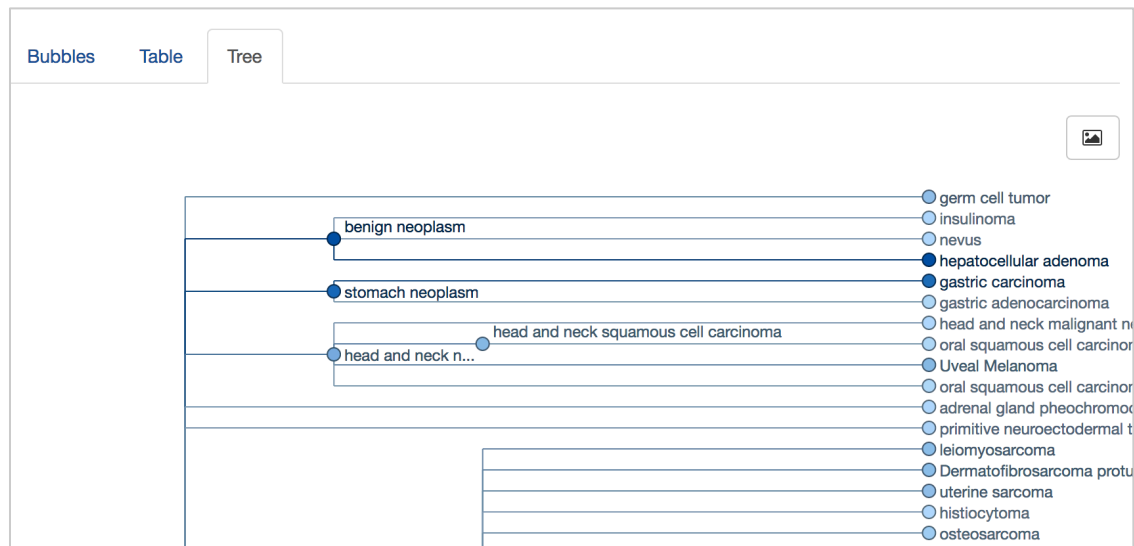
Search:

	Disease	Association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Therapeutic area
	neoplasm	1.00	1.00	0	1.00	1.00	0.23	0.10	0.30	
	cancer	1.00	1.00	0	1.00	1.00	0.23	0.09	0.29	neoplasm
	carcinoma	1.00	1.00	0	1.00		0.23	0.08	0.26	neoplasm
	renal carcinoma	1.00	1.00	0	0.78			0.06	0.09	neoplasm, kidney disease

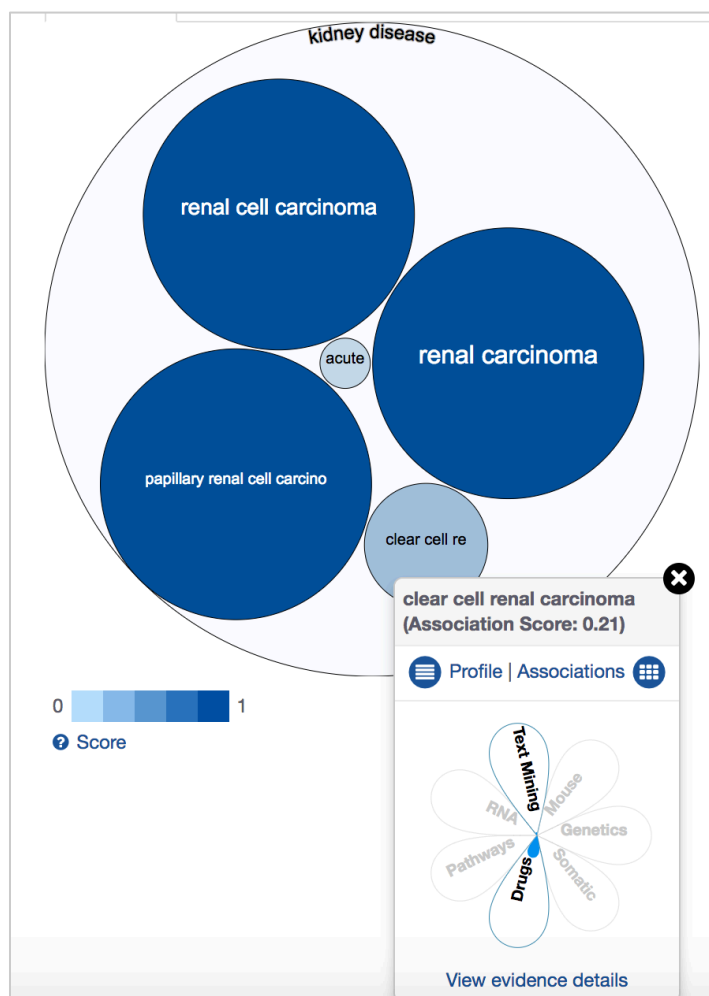
Tip: We colour code the cells in the table in different shade of blue as a visual way to convey the strength of the association (strongest association is coloured in dark blue). The score varies from 0 to 1. Hover over the cells in the table to view the numbers. Alternatively, you can select the cells 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 “Kidney diseases” associated with this target. There are five of them (including papillary renal cell carcinoma). The scores vary from 0.03 to 1.0.



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 prostate carcinoma when taking into account all evidence integrated in the Platform?

b) Can you restrict the results based on Somatic mutations only? Do you get the same genes (among the top 10) between this list and the one resulting from step (a) above?

Let's focus now on one of these targets namely *FGFR4* and find out more about evidence supporting the association with prostate carcinoma.

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) Click on the 'Browser' link to view the mutations is a graphical display. Are there variants associated with other traits (or diseases) in the region of the *FGFR4* gene?

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

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, the *MS4A1* gene encodes antigen CD20.

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) Continue with the same filters as in a). 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) Let's now explore some disease information available for non-Hodgkin's lymphoma. Can you list all drugs in phase IV of clinical trials for the treatment of this disease which status is completed (i.e. no longer recruiting patients for the clinical trials)?

e) Can you find the different subtypes i.e. the children terms of non-Hodgkin's lymphoma in its ontology? Can you download this image?

Note: you may want to click on the children diseases and see which targets have been associated with them.

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 auto-phosphorylation 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

1) the GTEx project

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 *EGFR*. Can you use the Open Targets Platform to find out if there is a mouse orthologue for this human gene?

f) Can you name the drug currently in clinical trial IV, status terminated, targeting this gene in non-small cell lung carcinoma?

g) OSIMERTINIB is another drug targeting *EGFR* also in clinical trials phase IV for many carcinomas. Does this drug target other kinases than EGFR? (Tip: you will be searching for this drug in the Open Targets Platform).

Demo 3

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

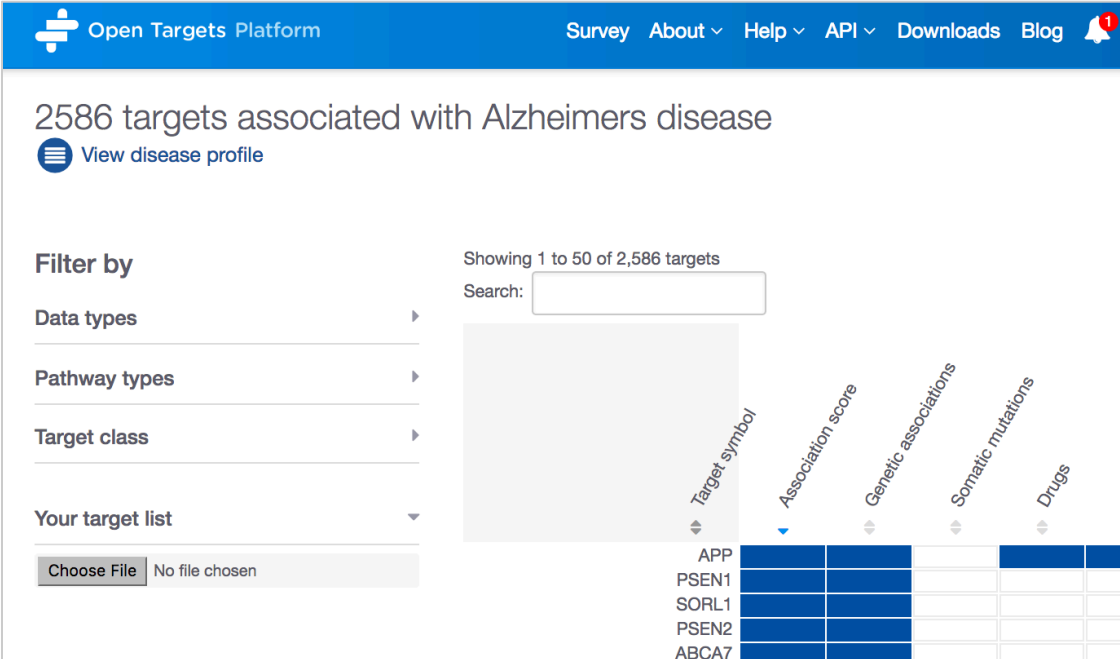
Franke et al (2006) listed a several genes associated with Alzheimer's based on a functional gene network:

ENSG00000091513
ENSG00000175899
ENSG00000143801
ENSG00000142192
ENSG00000130203
ENSG00000010704
ENSG00000080815

Let's see which information is available in the Open Targets Platform that could help you to choose and prioritise a target for follow up. We can input this data as a list (in .txt) to filter the associations in Open Targets.

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

Firstly, let's search for Alzheimer's and then upload our target list onto the Platform:



Open Targets Platform

Survey About Help API Downloads Blog

2586 targets associated with Alzheimers disease

[View disease profile](#)

Filter by

- Data types
- Pathway types
- Target class
- Your target list

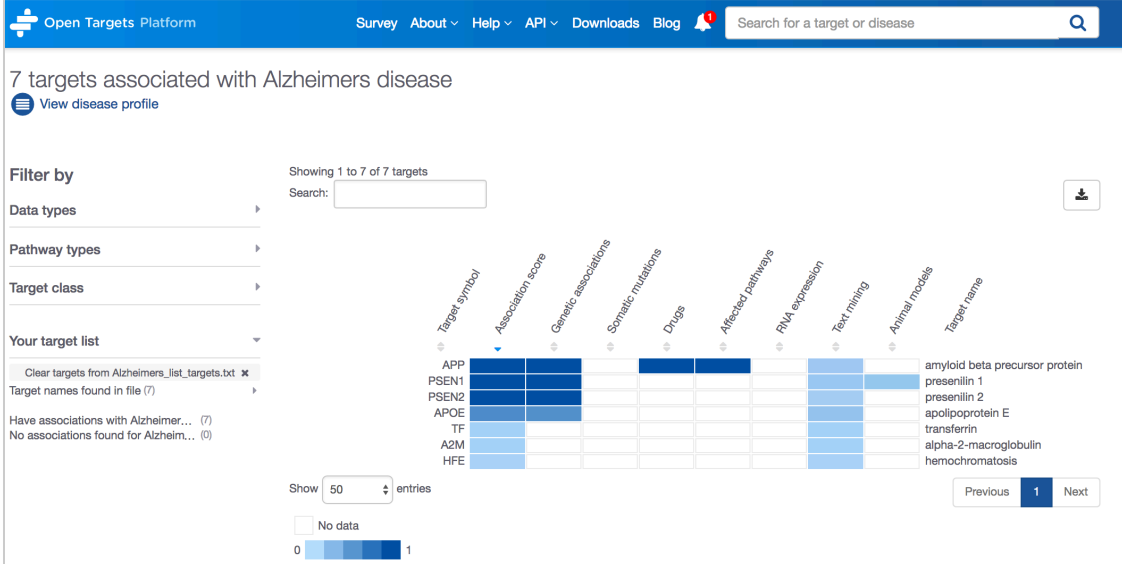
Showing 1 to 50 of 2,586 targets

Search:

Target symbol	Association score	Genetic associations	Somatic mutations	Drugs
APP				
PSEN1				
SORL1				
PSEN2				
ABCA7				

Choose File No file chosen

Now, we can upload a text file containing our list of genes (either as official gene symbols from HGNC e.g. *SOX3*, or Ensembl gene IDs e.g. ENSG00000134595). The filter 'Your target list' is at the left hand side of the association page (you will need to scroll down to see this option):




The screenshot shows the Open Targets Platform interface. On the left, under 'Filter by', the 'Your target list' filter is expanded, showing an option to 'Clear targets from Alzheimers_list_targets.txt'. A red arrow points to this section. The main area displays a heatmap for 7 targets associated with Alzheimer's disease. The targets are listed on the left: APP, PSEN1, PSEN2, APOE, TF, A2M, and HFE. The heatmap columns represent different data types: Target symbol, Association score, Genetic associations, Somatic mutations, Drugs, Affected pathways, RNA expression, Text mining, Animal models, and Target name. The heatmap shows various associations, with APP having the most associations. Below the heatmap, there is a 'Show 50 entries' dropdown and a legend for 'No data' (white) and '0' to '1' (blue gradient). The bottom right shows 'Previous', '1', and 'Next' navigation buttons.

The list should be uploaded and the resulting table will show the results of our analyses for that list of seven genes. This can help you prioritise which targets to follow up.

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

We can now use the filter 'Target class' to focus on membrane receptors only. There is only one, APP (ENSG00000142192)

1 target associated with Alzheimers disease

 [View disease profile](#)

Filter by

Data types ▶

Pathway types ▶

Target class ▼

Clear all ✕ Select all ✓

☐ Ion channel (2) ▶

☐ Secreted protein (1)

☒ Membrane receptor (1)

Your target list ▼

Clear targets from Alzheimers_list_targets.txt ✕

Target names found in file (7) ▶

Have associations with Alzheimer... (1)

No associations found for Alzheim... (6)

Showing 1 to 1 of 1 targets

Search:

Target symbol Association score Genetic association


APP

Show 50 entries

No data


0 1

Score



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

To explore more about the target itself, we can click on any cell of the resulting table then click on the target name, next to the flower:


 Open Targets Platform Survey

Evidence for APP in Alzheimers disease



APP
amyloid beta precursor protein
Synonyms: A4, AD1

Functions as a cell surface receptor, involved in neurite growth, neuronal development, and cell cycle regulation.



We will land on a page like this:

Open Targets Platform Survey About ▾ Help ▾ API ▾ Downloads Blog 1

APP

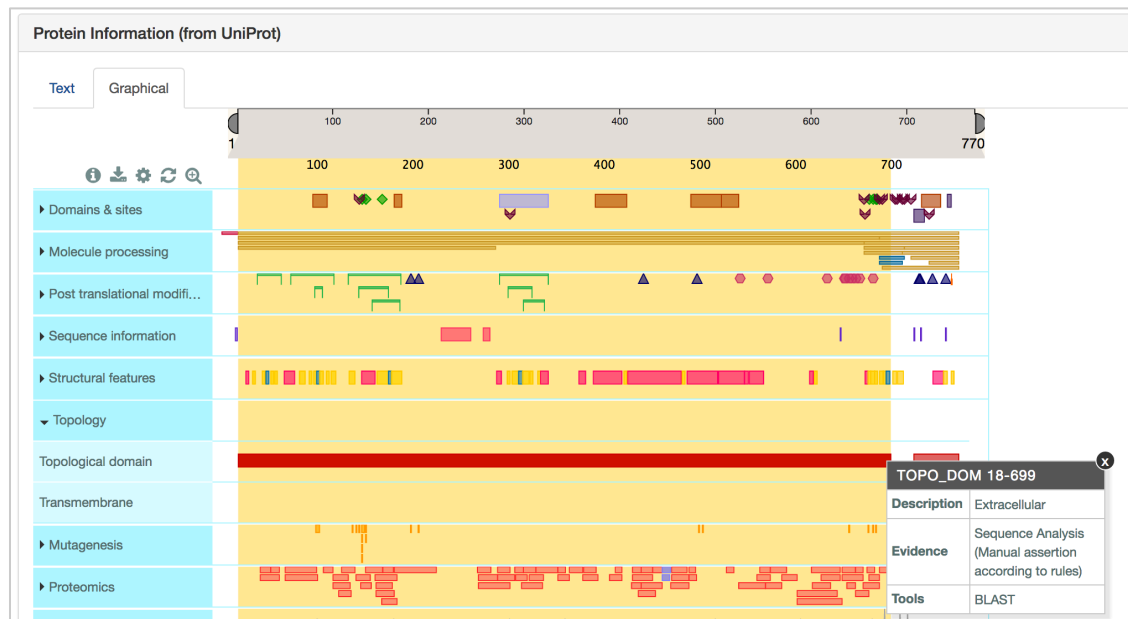
amyloid beta precursor protein | View associated diseases

Functions as a cell surface receptor and performs physiological functions on the surface of neurons relevant to neurite growth, cell motility and transcription regulation through protein-protein interactions. Can promote transcription activation through interaction with Numb. Couples to apoptosis-inducing pathways such as those mediated by G(O) and JIP. Inhibits G(o) and G(i) signaling. Membrane receptor, mediating the axonal transport of beta-secretase and presenilin 1. Involved in copper homeostasis. Metallated APP induces neuronal death directly or is potentiated through Cu(2+)-mediated low-density lipoprotein oxidation. Components of the extracellular matrix such as heparin and collagen I and IV. The splice isoforms that contain ... [\[show more\]](#)

Synonyms: A4 AD1 CVAP Amyloid beta A4 protein PN-II peptidase nexin-II Cerebral vascular amyloid peptide Beta-amyloid protein
APPI PreA4 Alzheimer disease amyloid protein APP

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

To find what which amino acids of this membrane receptor correspond to the extracellular domain, we need to expand the tab 'Protein Information (from UniProt)', click on 'Graphical', then expand the 'Topology' section in the image:



Click on the red boxes in 'Topology' to get a pop up window with more information. You will find out the extracellular domain is between amino acid 18 and 699.

EXTRA HANDS-ON EXERCISES

If you have finished exercises 1, 2 and 3 above, you may want to try these too:

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 increased 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/>