Open Targets: integrating genetics, omics and chemical data for drug discovery

7th RoBioinfo Seminar

Cluj-Napoca

8th October 2019

Denise Carvalho-Silva, PhD

EMBL-EBI | Open Targets
United Kingdom

Welcome to our training session!

This booklet is based on the September 2019 release (19.09) of the Open Targets Platform. Check the links below for more information:

Open Targets Platform:

Release blog posts
http://blog.opentargets.org/tag/release-notes/

Help documentation
https://docs.targetvalidation.org/

FAQs
https://docs.targetvalidation.org/faq/frequently-asked-questions

Open Targets Platform paper
http://bit.ly/cite-us

Open Targets Genetics:

Release blog posts
http://blog.opentargets.org/tag/genetics-releases/

Help documentation
https://genetics-docs.opentargets.org/

FAQs
https://genetics-docs.opentargets.org/faqs

Questions or suggestions?
support@targetvalidation.org
geneticsportal@opentargets.org

TABLE OF CONTENTS

OVERVIEW

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

We work to create a research and development (R&D) framework that can be applied to a wide range of human diseases. We share our results 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 Sanger Institute (http://www.sanger.ac.uk/) and the EMBL-EBI (European Bioinformatics Institute) (http://www.ebi.ac.uk/). In February 2016, Biogen (https://www.biogen.com/) joined the initiative. The consortium was rebranded to Open Targets in April 2016, and has welcome two three new partners since, namely Takeda in 2017), and Celgene and Sanofi, both in 2018.

In drug discovery, the *validation* of a target refers to the creation of a specific entity that modulates the activity of a target to provide therapeutic benefit to individuals with a disease.

The ultimate validation of a target is the creation of an effective therapeutic molecule. This is a long and costly endeavour with more high failure rates.

The goal of Open Targets is to transform this process by predicting if the modulation of a target is likely to provide therapeutic benefit. This would be done much earlier in the drug discovery process than is currently possible and far in advance of having a final, approved medicine.

Points covered in this workshop:

- The science carried out in Open Targets
- Introduction to the Open Targets Platform
- How to make the most of the Open Targets Platform
- Overview of alternative ways to access the Open Targets Platform data

INTRODUCTION TO OPEN TARGETS

Open Targets employs large-scale human genetics and genomics data to change the way drug targets are identified and prioritise. We have established a set of scientific projects to both **integrate** and **generate** data and analytical processes that implicate a target as valid.

Our experimental projects use CRISPR gene editing, induced pluripotent stem cells, single cell genomics, organoids to generate new data and provide insights in the validation of targets relevant to key therapeutic areas namely:

- Oncology
- Immunology
- Neurodegeneration

The results of these experimental projects have been published in Nature (PMID:30971826), Nature Medicine (PMID:31209336), or are available on bioRxiv (https://www.biorxiv.org/content/10.1101/654632v1).

Our informatics teams have developed several resources for drug discovery, notably the Open Targets Platform, a tool that provides easy access to data relevant in drug target identification and prioritisation. More recently, we have also launched Open Targets Genetics for the exploration of variant-gene-trait associations from UK Biobank and GWAS Catalog studies.

For more details on our projects, go to Scientific Overview.

Open Targets Platform

The Open Targets Platform is a web application that integrates, scores and displays publicly available data to facilitate the identification and selection of targets for new therapies.

We use genetics, omics and chemical data from different data sources to associate genes and diseases. Similar data sources are combined into the following data types:

Genetic associations
Somatic mutations
Drugs
Pathways & systems biology
RNA expression
Text mining
Animal models

The evidence (e.g. SNPs, scientific literature) is used to compute the association score, a four-tier framework that assesses the frequency of the evidence, the confidence and severity (e.g. does the SNP change the amino acid of the target protein?) and aggregates evidence scores using the sum of the harmonic progression. The resulting aggregation gives rise to the data source, data type and overall scores. The association score can be used to rank target and disease associations in the Platform. The latest release of the Platform (September 2019) contains:

- 27,024 targets
- 10,474 diseases
- 3,336,588 associations between targets and diseases
- 7,784,471 evidence points

The Open Targets Platform is an open source and open access tool that can be applied in a variety of use cases in academia and pharmaceutical companies. We kindly ask you to cite our latest paper (PMID:30462303) if you use our data in our work.

What can you do with the Open Targets Platform?

- Search for a disease (and target) and find its associations based on genetic and omics data, drug information, text mining, etc
- Delve deep into the evidence supporting these associations
- Prioritise targets based on annotations at the target level e.g. tractability and safety data

- Find annotations for diseases, such as all drugs marketed or approved for clinical trials
- Search for several targets at once with the batch search
- Carry out more complex queries using the REST API
- Download all evidence and association objects for downstream analyses

Connect with us

❖ Open Targets Blog

❖ Follow us on Twitter, Facebook, LinkedIn, and YouTube

OPEN TARGETS PLATFORM: WALKTHROUGH

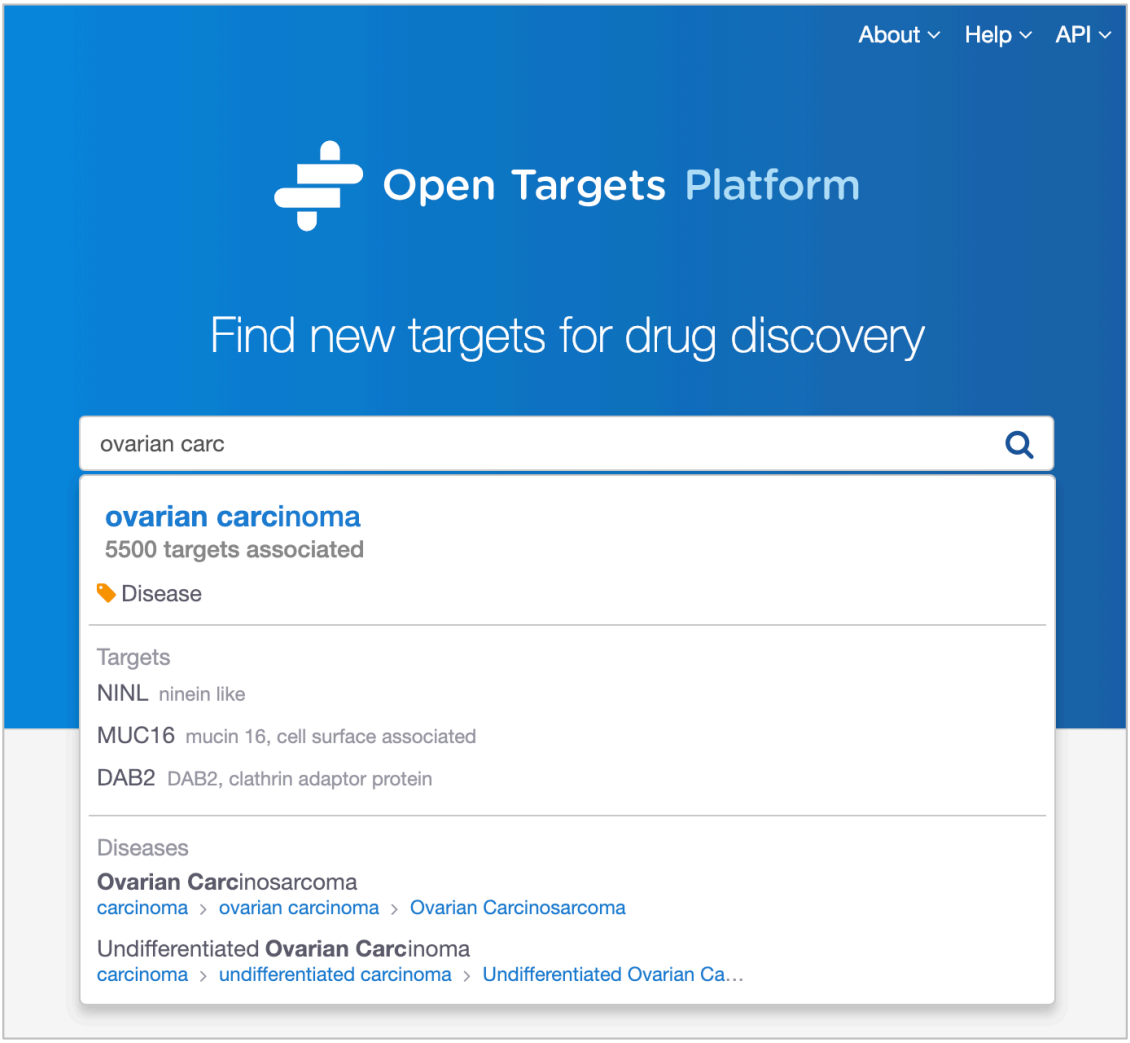
We will guide you through the website using ovarian carcinoma, as an example of a disease, then we will explore the evidence associating ERBB2 with that disease.

The following points will be addressed during the walkthrough:

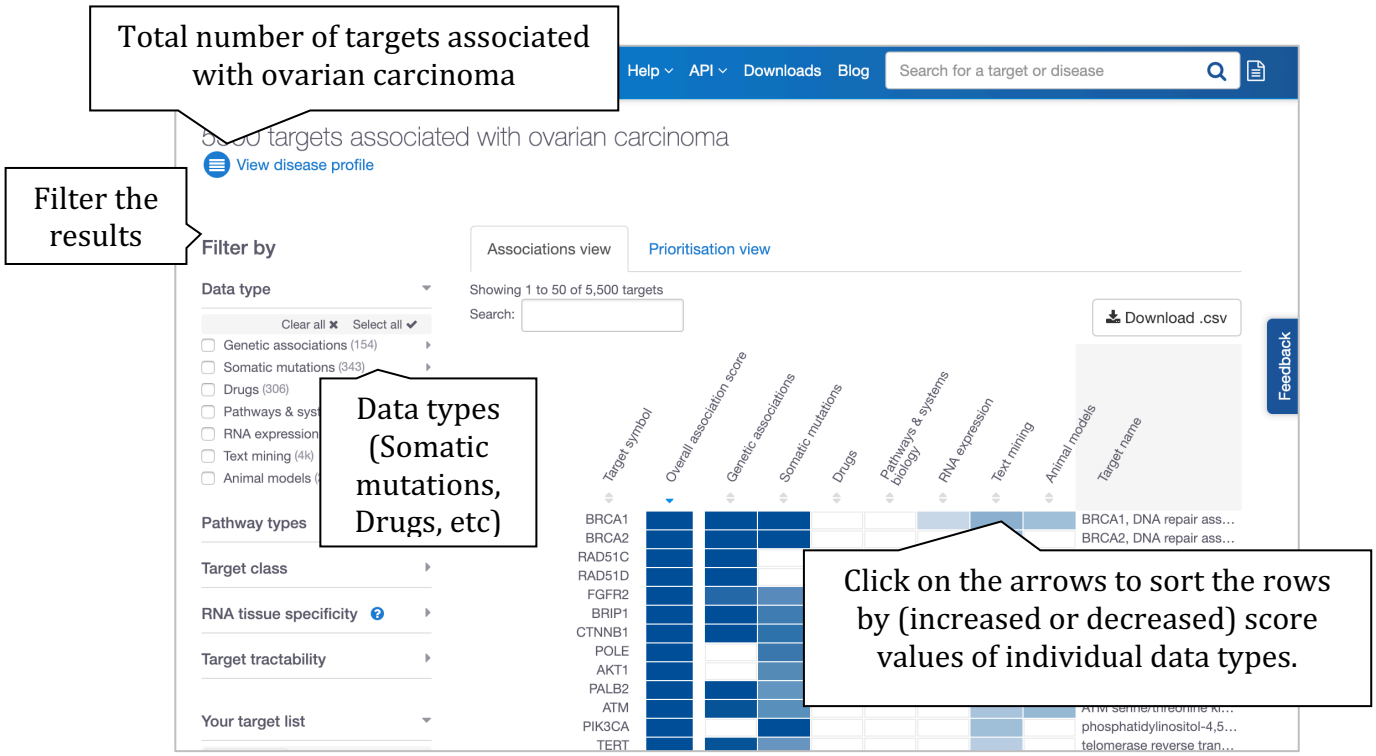
- How to find targets associated with ovarian carcinoma
- How to filter down the number of targets based on specific types of evidence, pathway types, target class, etc
- How to find out how strong the association between ERBB2 and ovarian carcinoma is
- How to view the evidence that supports this association
- How to find other diseases associated with ERBB2
- How to visualise the ERBB2 gene and its variants on the genome
- How to find drugs currently in clinical trials for ERBB2
- How to filter the target associations for ovarian carcinoma by using a list of genes

Demo 1: Searching for a disease

Go to www.targetvalidation.org:



Search for ovarian carcinoma and select the first hit to be directed to a page like this:



The current release of the Open Targets Platform (September 2019) lists 5551 targets associated with ovarian carcinoma.

The above table is sorted by default with the best hit at the top of the table. This first target is the gene that contains the highest number of supporting evidence points. This is summarised by the overall association score, varying from 0 to 1 (the closer to 1, the more evidence Open Targets Platform has for an association).

The association score is computed in four steps:

1) Firstly, we compute an **evidence score** for each piece of evidence that is used to support an association e.g. one SNP
2) Then we compute the **data source score** for all evidence within a data source e.g. all SNPs from GWAS Catalog supporting a target-disease association
3) The third step is when we aggregate all data source scores to compute the **data type score**

The **Overall association score** is the aggregation of all data source scores.

Note 1: whenever we aggregate scores, we apply the harmonic sum:

$Score = S_1 + S_2/2^2 + S_3/3^2 + S_4/4^2 + S_i/i^2$

Note 2: The ranking in the image above (e.g. BRCA1 > BRCA2 > RAD51C) relies on the number of evidence available for the individual associations. Although the overall association score is 1 for all of those three targets, we have more evidence available for BRCA1; hence BRCA1 comes before BRCA2 and RAD51C.

We also apply different weights to different data types whilst computing our score. RNA expression, animal models and text mining data are all down weighted by a factor of 0.2. Sysbio, PROGENy and SLAPenrich are also down weighted by a factor of 0.5.

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

Check our help documentation pages to find out more about Open Targets Platform data sources or association score.

Let's go back to the association table shown above. It lists all targets associated with ovarian carcinoma. These associations can be filtered by six categories:

1) Data types
2) Pathway types
3) Target class
4) RNA tissue specificity
5) Target tractability
6) Your target list

Let's have a look at the individual options for each of these filters:

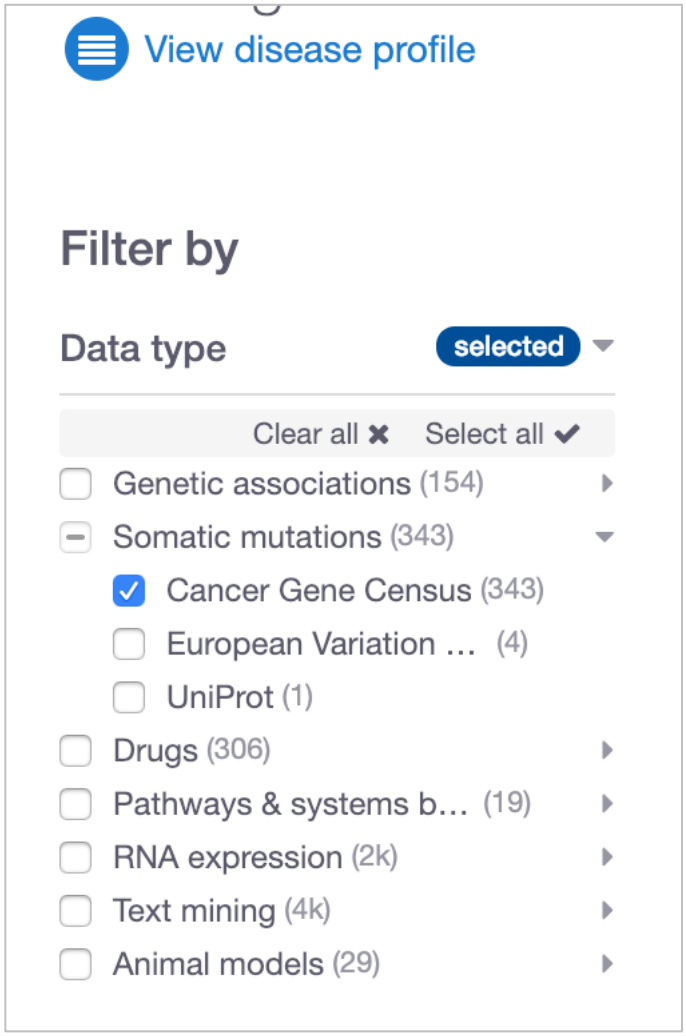
1) Data types: we collect data from various sources and combine them into categories called Data types. Examples of data sources are GWAS catalog and UniProt, both combined into the Genetic associations data types. Note that data from an individual source can

contribute to different data types, e.g. data from EVA is observed in two data types, Genetic associations and Somatic mutations.

These are the data types where evidence was used for the associations with ovarian carcinoma:

Genetic associations
Somatic mutations
Drugs
Pathways & systems biology
RNA expression
Text mining
Animal models

Note: as our data types above can be made up of different data sources, click on the grey arrows to expand the options and filter the table to focus on specific data sources, such as Cancer Gene Census (data source) within Somatic mutations (data type):



2) Pathway types: Reactome is the resource that provides us with pathway classification e.g. immune system (and its subtypes e.g. neutrophil degranulation), hemostasis (platelet degranulation), etc.

These are some of the pathway types currently available:

Immune System
Signal Transduction
Metabolism
…

3) Target class: ChEMBL provides us targets into grouped into different classes such as Enzyme, Ion channel, Membrane receptor, etc.

4) RNA tissue specificity: RNA tissue specificity: the tissue specificity of a target is computed as the number of standard deviations from the mean of the log RNA expression of the target across the available tissues. This is a standard z-score calculation. A target is considered to be tissue specific if the z-score is greater than 0.674 (or the 75th percentile of a perfect normal distribution). We remove data for under-expressed targets before the z-score calculation. This RNA expression data comes from Expression Atlas.

This filter allows users to select the organs (or anatomical system) where the target is significantly more expressed in the selected tissues than the mean of the other tissues.

5) Target tractability: tractability of a target is the confidence that we can identify a modulator that interacts with the target to elicit a desired biological effect. We currently assess whether the targets can be modulated by two drug modalities: Small molecules (SM) and Antibodies (Ab). This data is a modified version of the tractability method described by Brown et al. 2018, who have assigned tractability assessment to different buckets, ranging from 1 to 8 (or 9 for Ab), 1 being the highest degree of tractability.

In the Open Targets Platform, the different buckets for SM tractability are grouped into three categories:

- Clinical precedence: if drugs are already in clinical trials

Buckets 1, 2 and 3

- Discovery precedence: if there are active compounds in ChEMBL or targets with ligands in PDBe (Protein Data Bank in Europe)

  Buckets 4 and 7

- Predicted tractable: if some tractability metrics such as drugEBIlity and/or 'Ro5 druggable' domain

- Buckets 5, 6 and 8

For antibodies, the bucket assignment (1-9) is grouped into the following three categories:

- Clinical precedence: if drugs are already in clinical trials

  Buckets 1, 2 and 3

- Predicted tractable (high confidence): if the location of the protein is known (from UniProt) or a cellular component is annotated by the Gene Ontology (GO)

  Buckets 4, 5 and 6

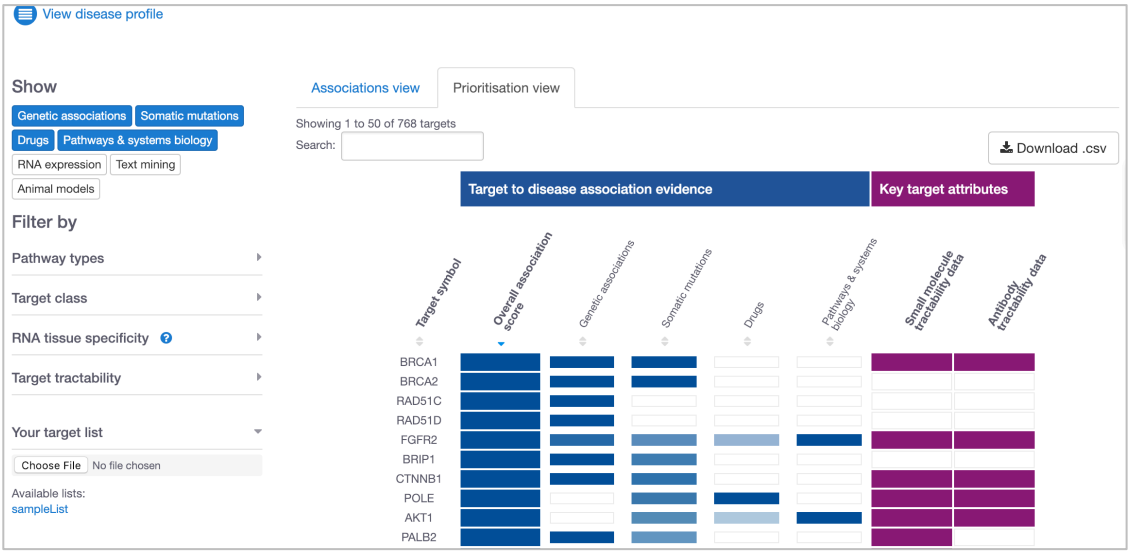
- Predicted tractable (mid to low confidence): if the location of the protein is unknown (from UniProt) and/or if a signal peptide and TM domain can be predicted or if there is information on the protein from the Human Protein Atlas

  Buckets 7, 8 and 9

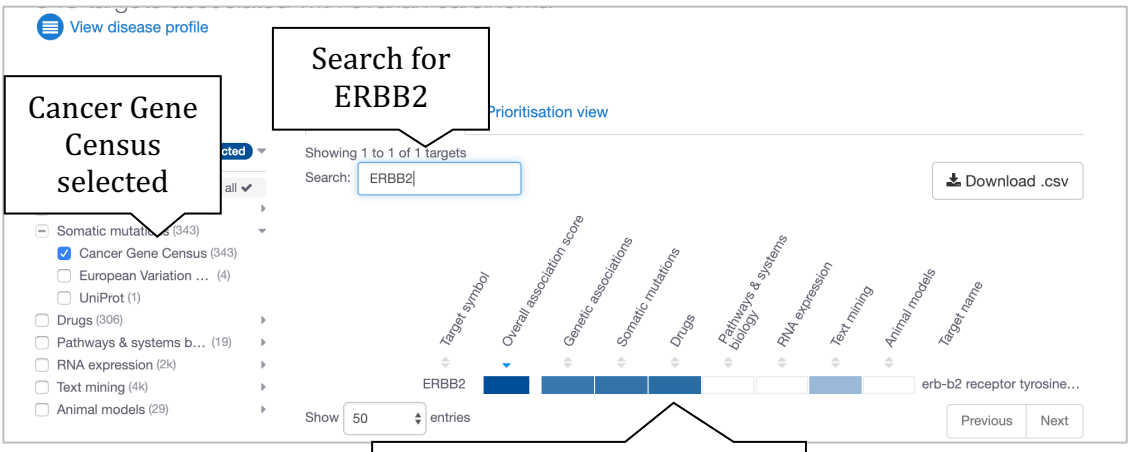
7) Your target list: in the associations page for a given disease, you can also filter the targets based on your own targets of interest. You can upload a list of targets (as .csv or .txt) and restrict the table to show only the targets in your list. This can help you to see the evidence Open Targets has integrated for your targets. Your own list of genes should be noted in official gene symbols from HGNC or Ensembl Gene IDs.

Note 1: "Your target list" is a slightly different functionality than the "batch search" tool. To use the former, you need to be at the associations page for a given disease, such as asthma, and filter all the associations down to the ones in your file. The batch search tool, on the other hand, will allow you to search for associations across the entire set of diseases in the Open Targets Platform.

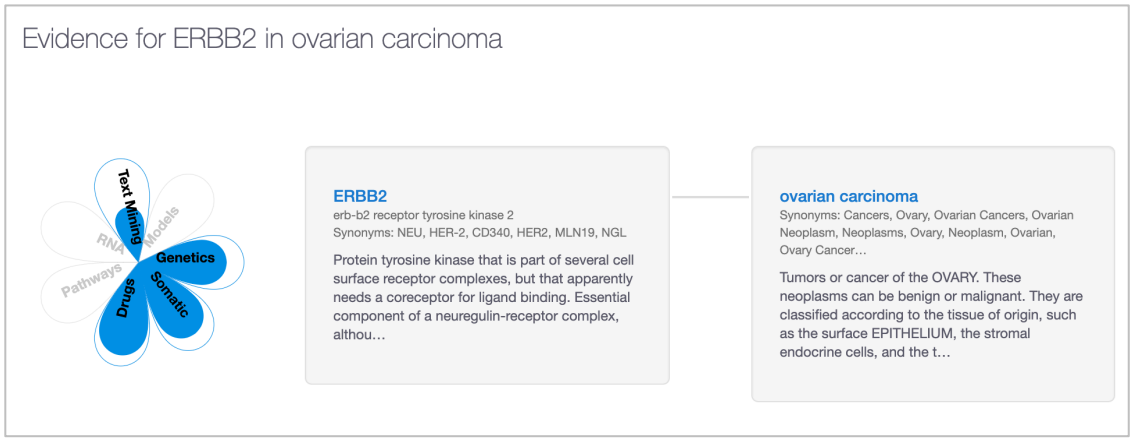
Note 2: In the associations page, there are two views that you can explore: Associations view and Prioritisation view. The latter shows whether the available targets can be modulated by small molecule or antibody.



Now that we have looked at the different filters and explored the Prioritisation view, let's now restrict our data based on evidence coming from Cancer Gene Census only. The number of targets goes down to 344. Let's focus on ERBB2 by searching for this gene symbol:



Click on any of the blue cells in the table above to go to the evidence page for the association between ERBB2 and ovarian carcinoma:



In the evidence page, you can explore the underlying data used for the association.

The coloured petals in the flower plot represent the data types that support this association. Grey areas in the flower plot indicate there is no information for the corresponding data types. This may change in future releases when more data becomes available.

Let's now scroll down on the page and expand the tabs available:

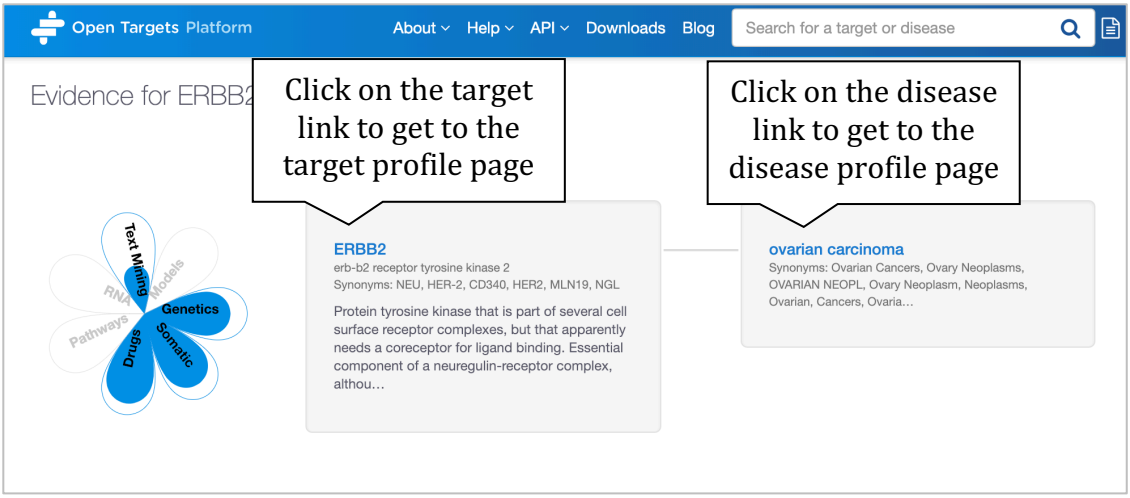
- Genetic associations with two tabs (Table and Browser)
- Somatic mutations (with links to COSMIC)
- Drugs (with links to clinicaltrials.gov)
- Text mining (with links to the original publications)

There are currently four **unique** drugs in different phases of clinical trials targeting ERBB2 in patients with ovarian carcinoma. Note the number of entries in the drugs table: 9 entries (evidence) for 4 unique drugs.

Now, let's have a look at the results from our text mining approach and find which papers have been used as evidence for the associations between ERRB2 and ovarian carcinoma. During the text mining, we are looking for the co-occurrence between the target name (or its synonyms) and the disease name (or its synonyms).

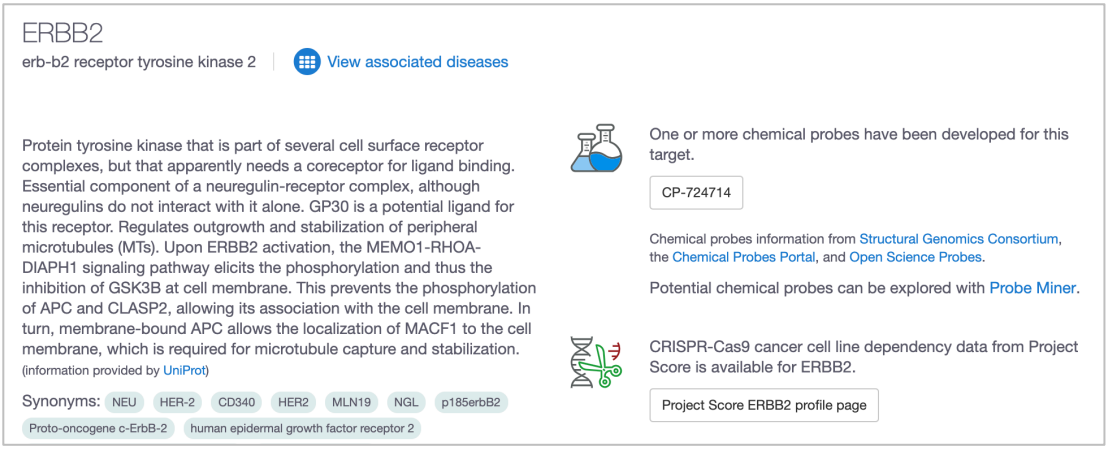
Demo 2: Target and disease annotations

Let's now scroll back up to the top of the evidence page and click on the "ERBB2" link:

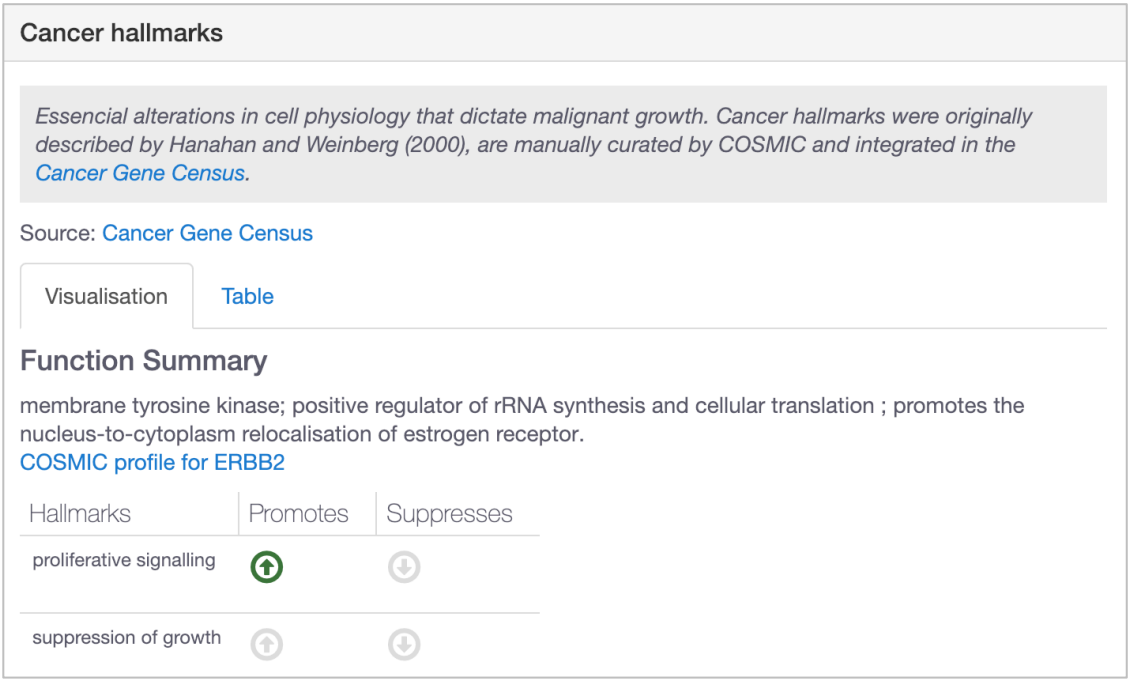


When clicking on ERRB2 in the above image you will be redirected to the Target profile page of this target:

https://www.targetvalidation.org/target/ENSG00000141736



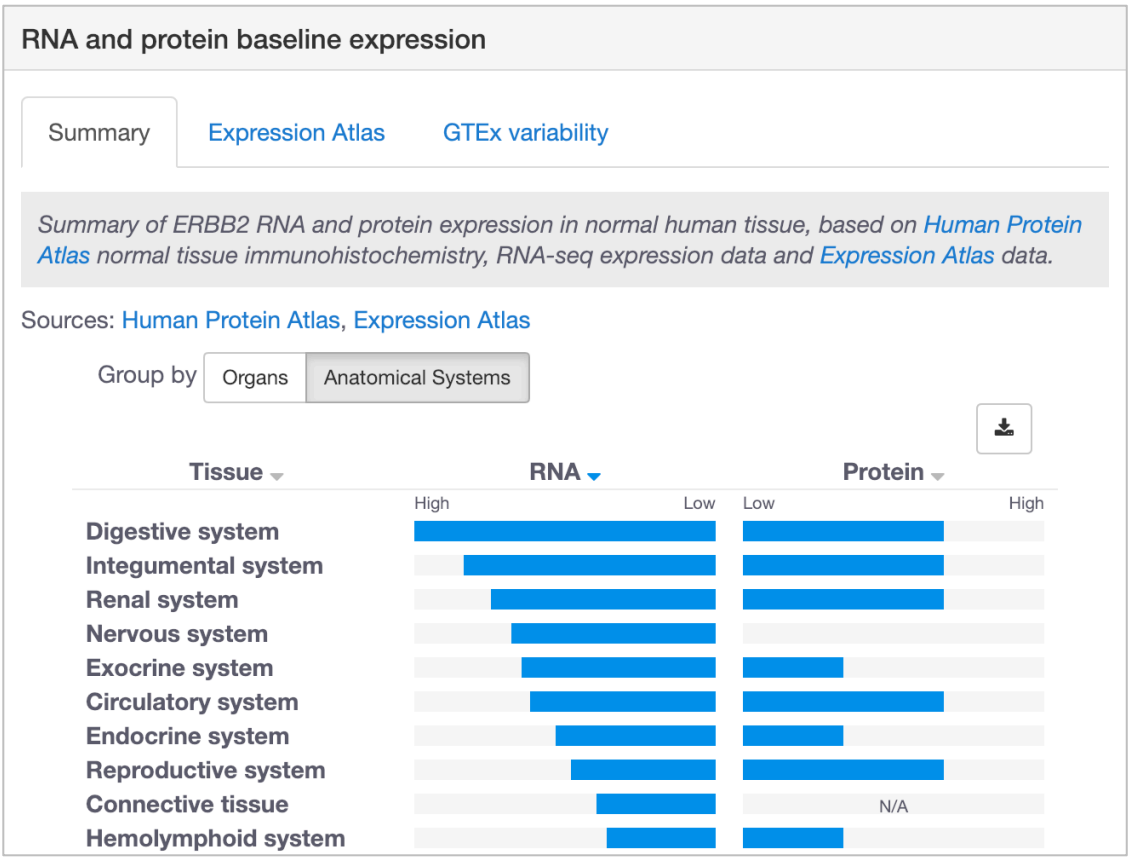
In the target profile page, you can find gene specific annotations that can help you to prioritise this gene as a drug target, namely RNA and protein baseline expression levels, protein structure, gene ontology terms, information on tractability of ERRB2, cancer biomarkers, target safety information, cancer hallmarks among many other annotations.



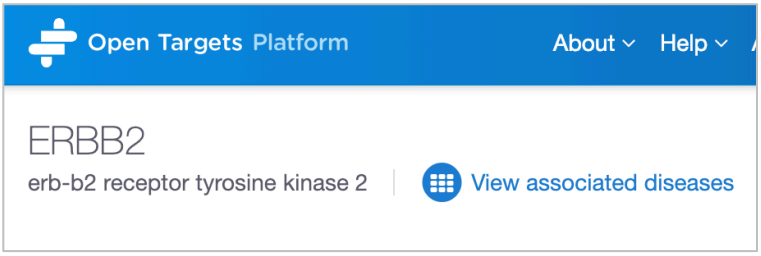
Let's now expand the RNA and protein baseline expression to find out in which organs or anatomical systems ERRB2 is expressed.

You will find three tabs in there: Summary, Expression Atlas (data from several projects including the Illumina Body Map) and GTEx variability.

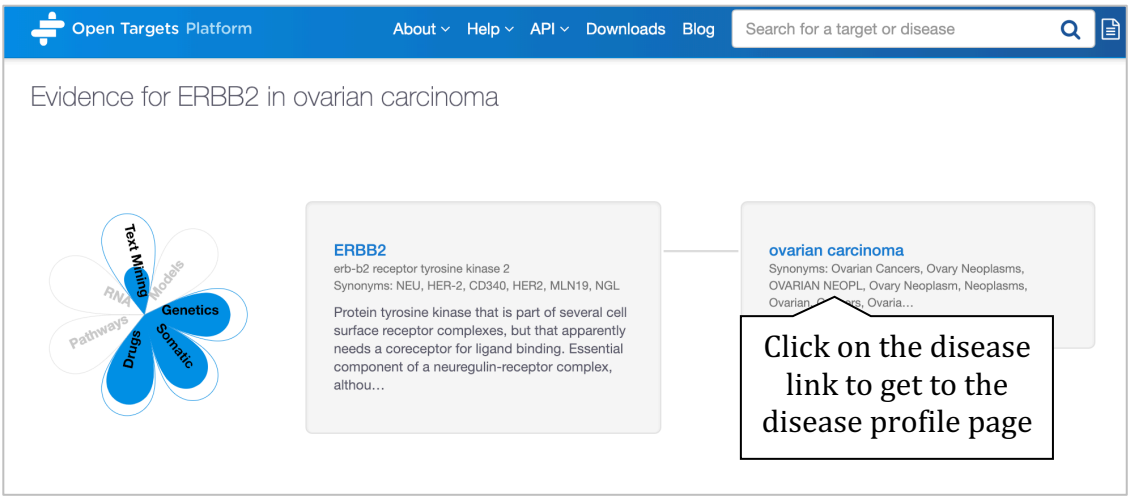
In the Summary tab, you can compare the mRNA and protein expression side by side. You will learn that the expression in the esophagus is higher at the protein level than RNA level. You can click on the tissue names to get further details on specific sections of the tissue/organ.



Note: In the target profile page, you can also explore other diseases associated with ERBB2, apart from ovarian carcinoma. Click on "View associated diseases" to see the associations for ERBB2.

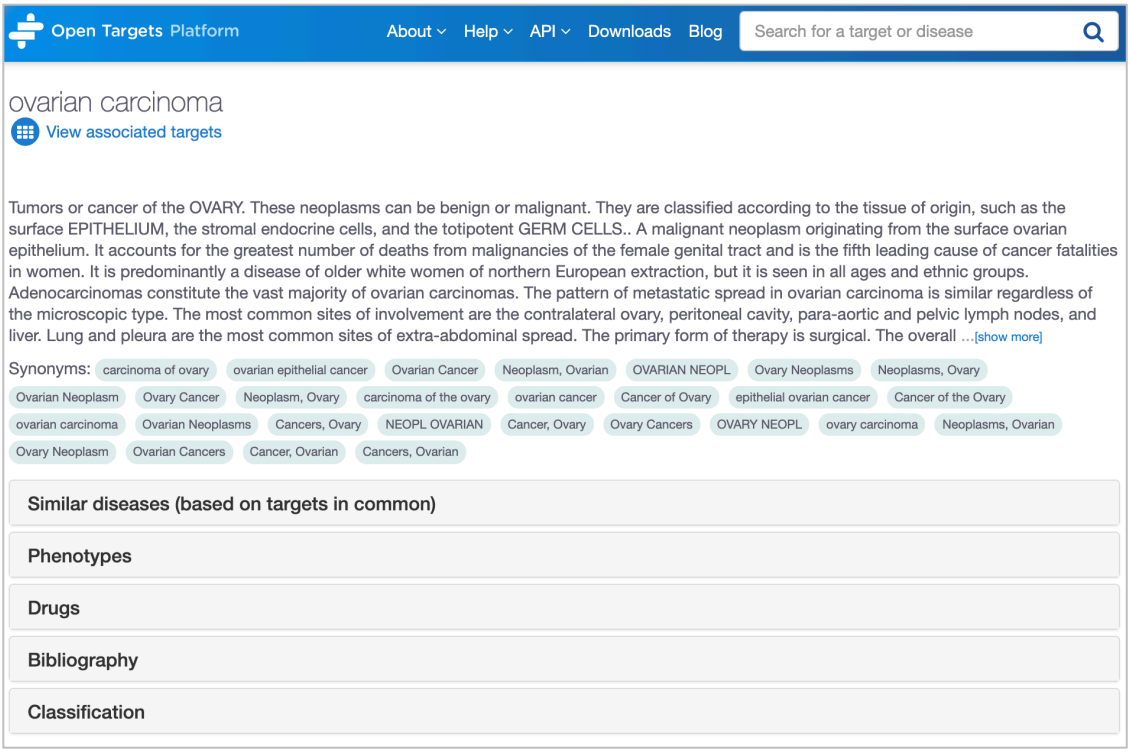


Let's now go back to the previous evidence page (the flower page):



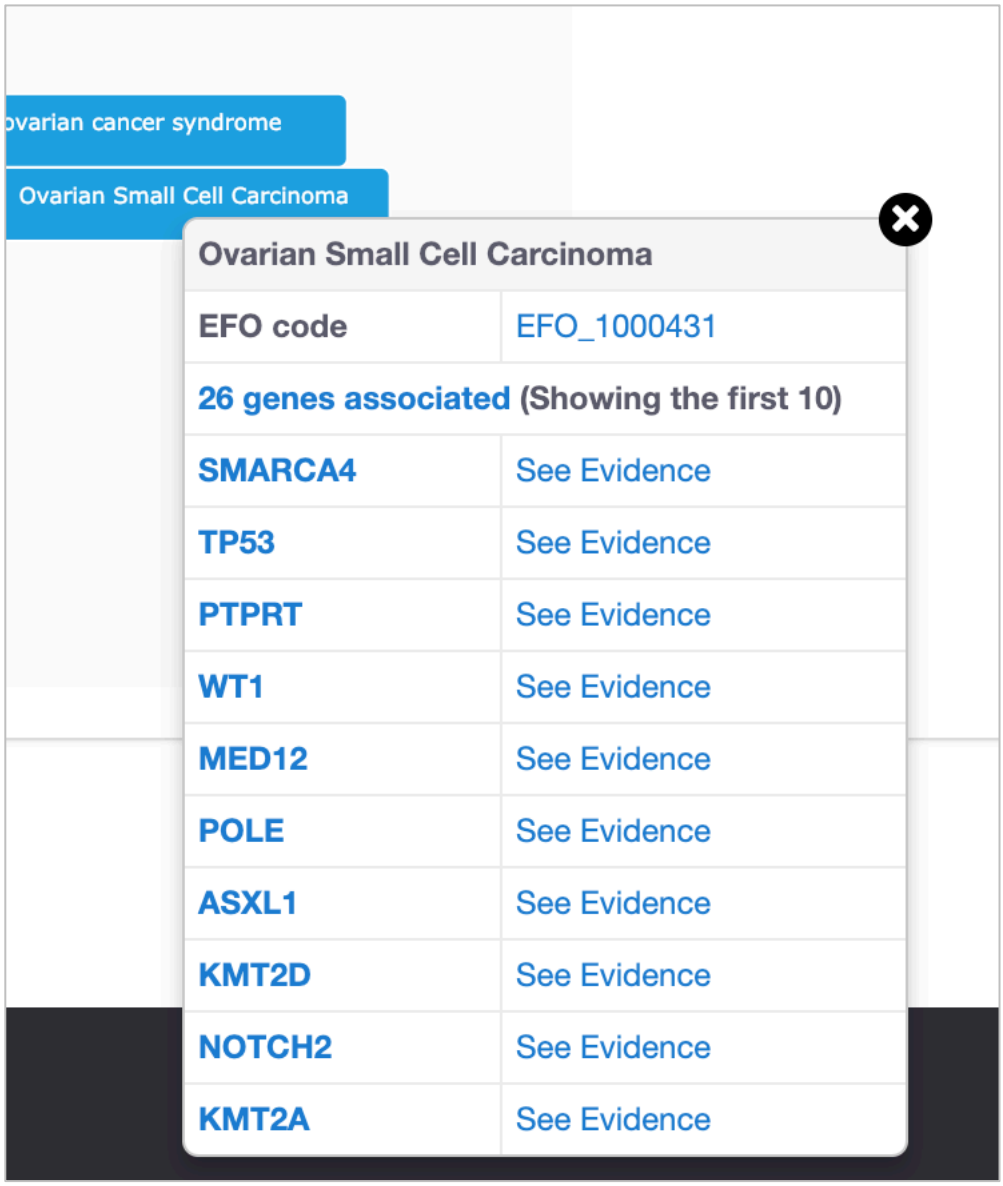
We can now click on the disease name and explore the annotations for ovarian carcinoma:

https://www.targetvalidation.org/disease/EFO_0001075



Under the Classification tab, you can see the disease ontology (disease concepts and relationships) from the EFO (Experimental Factor Ontology), an ontology developed and maintained by EMBL-EBI.

Ovarian carcinoma is represented in green, red nodes correspond to parental terms, whereas the children terms of ovarian carcinoma are shown in blue (e.g. ovarian small cell carcinoma). Click on any of the disease names to get a pop-up box with the (first) 10 targets associated with that disease:

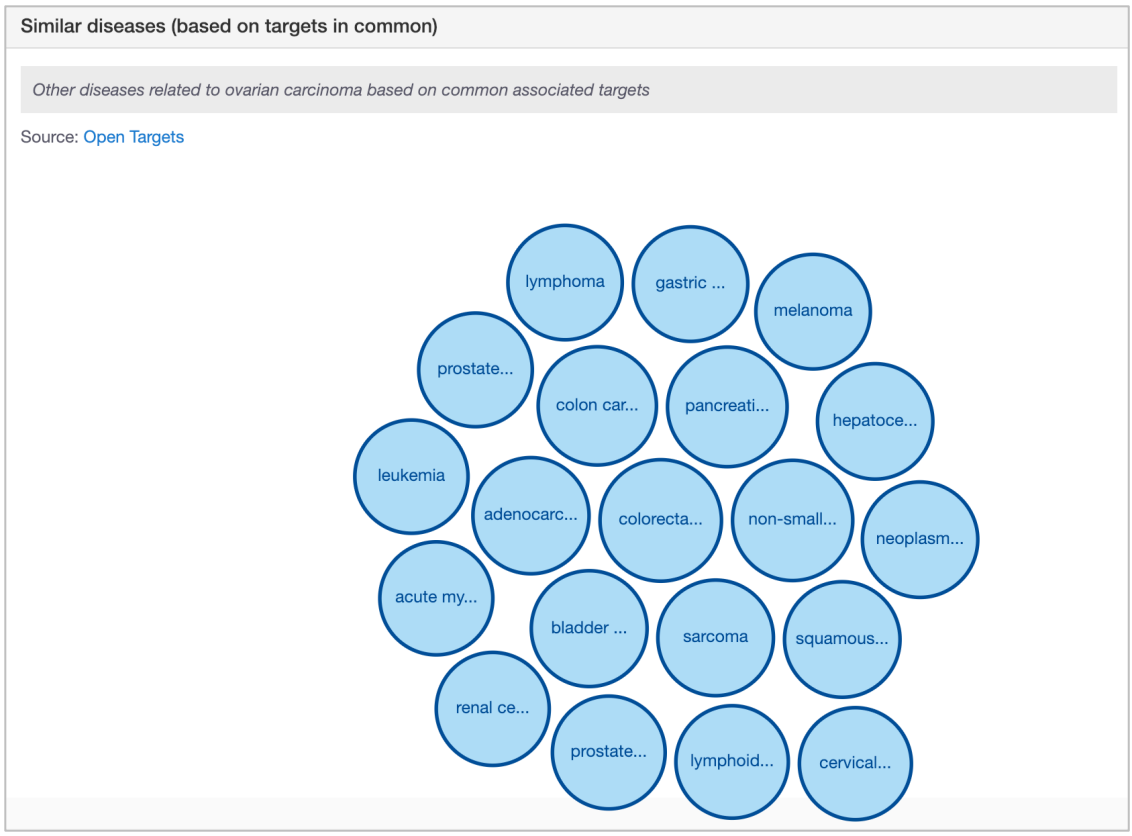


By relying on the disease ontology and the relationships it reveals, we can derive new associations that do not have direct evidence. For instance, IBD is an autoimmune disease that will have direct evidence for its association with its targets. We can propagate this direct evidence up to higher terms in the ontology of IBD and use this evidence (now indirect) to associate target X with autoimmune disease (a parent term of IBD). This procedure can allow us to find common targets across groups of related diseases (e.g. Ulcerative

Colitis, Crohn's disease and IBD) even when direct evidence is not available.

For more on this, check our blog post: https://blog.opentargets.org/direct-versus-indirect-evidence-should-you-care/

Still on the disease profile page, check the Similar diseases (based on targets in common) tab:

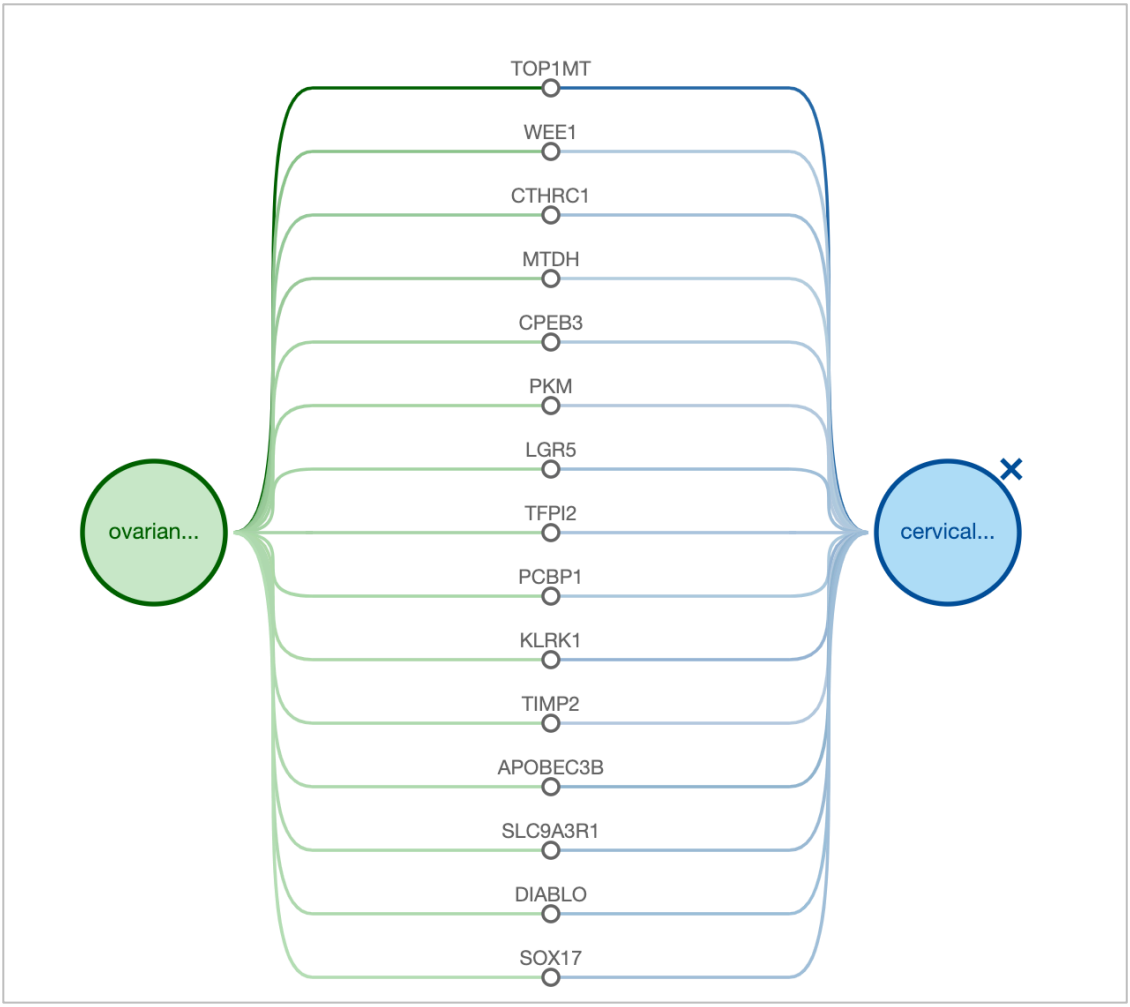


For each pair of diseases, we compute the overlap of shared targets against the total number of connections to both targets, correcting each pair by the significance and the specificity of these connections.

This procedure will consider targets that are specifically linked to fewer diseases more relevant than targets that are commonly linked to many types of diseases. More details on the framework for the Similar diseases feature can be found in our latest publication: http://europepmc.org/abstract/MED/30462303.

Note: You will see that the target profile page has a similar visualisation under the tab Similar targets (based on diseases in common). There we will compute a closer distance between two targets sharing a rare disease than two targets sharing diseases that are highly connected to many genes, such as cancer.

Click on any of the bubbles to get details on the targets in common between any two diseases and the evidence used for the associations (conversely for the diseases in common among any two targets that you can see in the target profile page):



Demo 3: Searching for a target

You can also use the Open Targets Platform if you want to start your search using a target, rather than a disease. Let's search for AR to explore visualisations and other functionalities available in our user interface.

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

- Table view

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

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

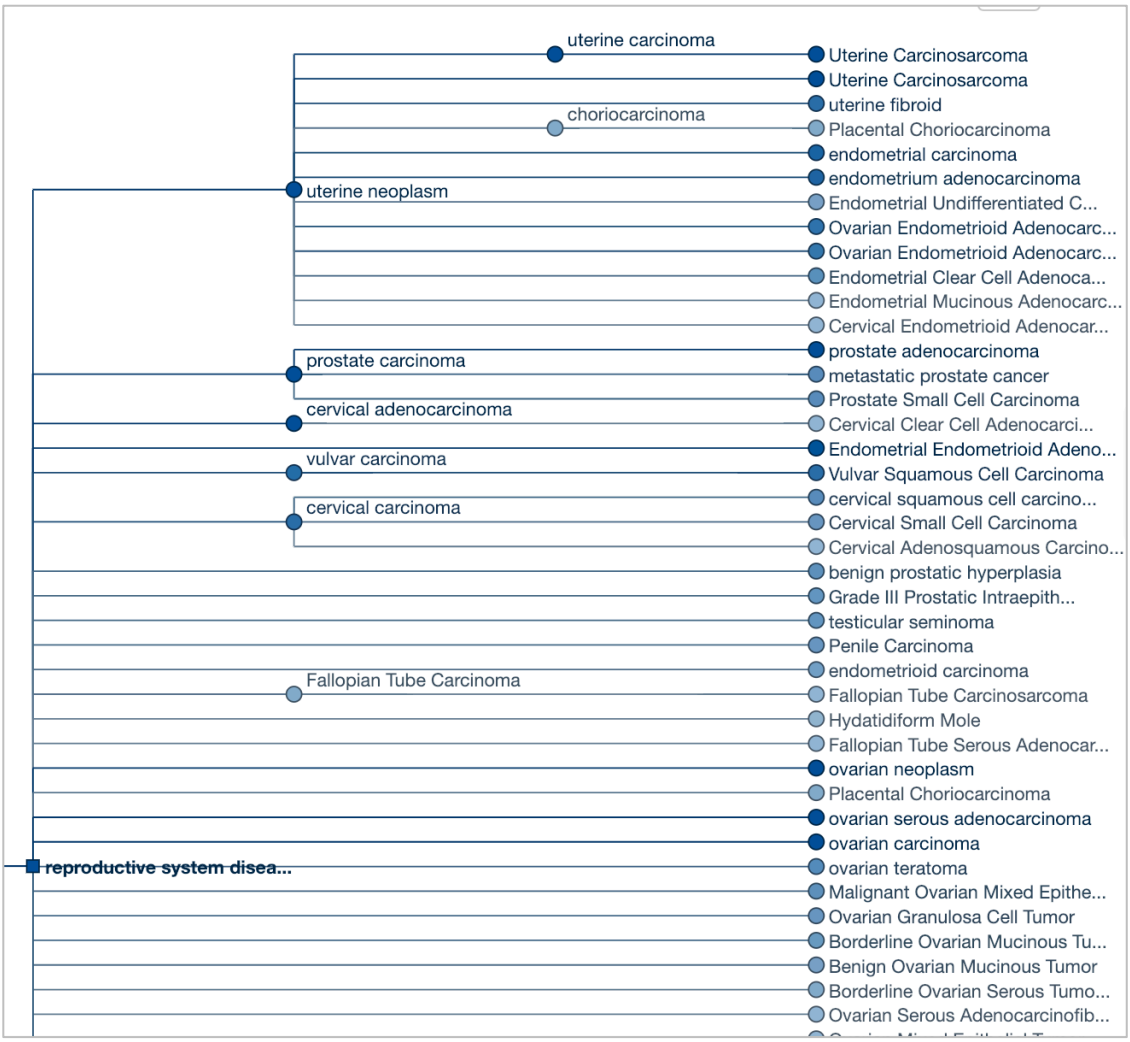
Tip: The different shades of blue in the table convey the strength of the association based on the available evidence (strongest association is represented in dark blue). The score varies from 0 to 1. Hover over the cells in the table to view the numbers. Alternatively, you can select the cells in the table so that you can view the numerical values.

- Bubbles 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.

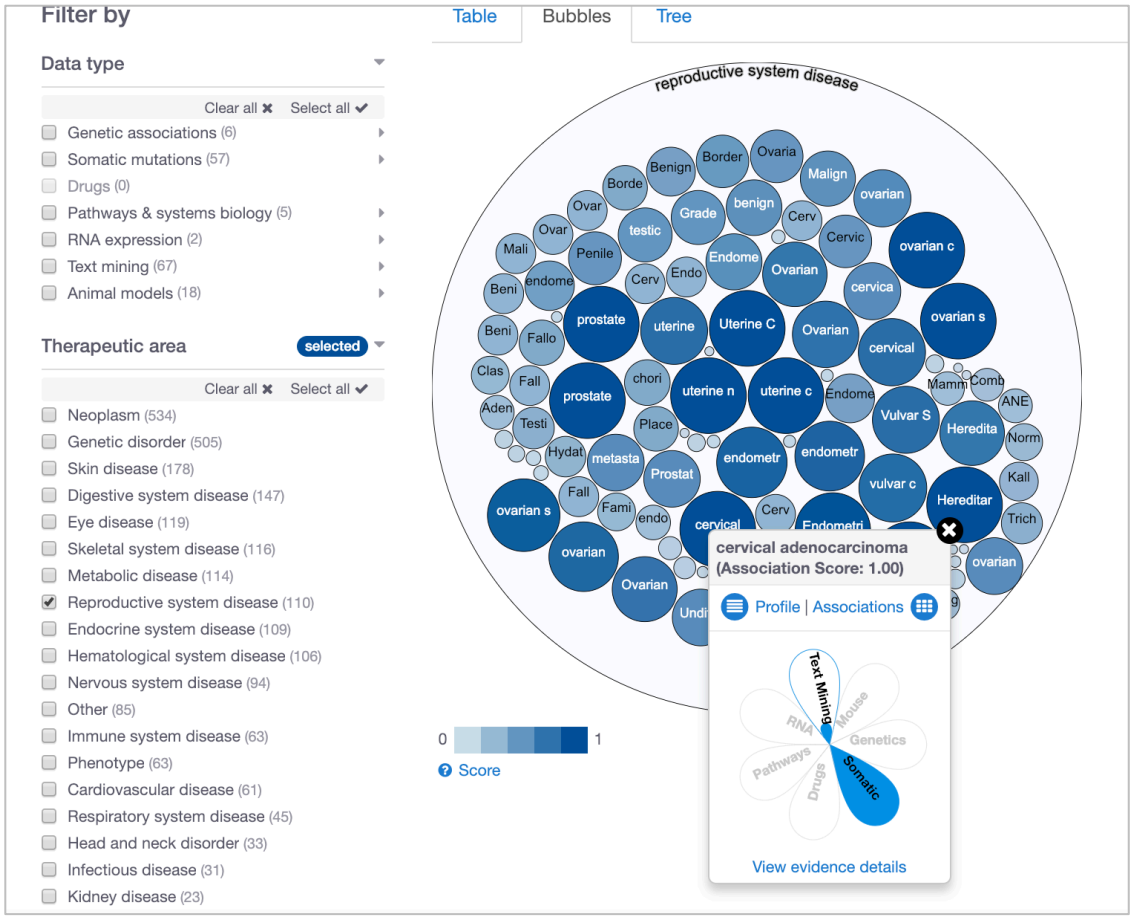
- Tree view

In the Tree view, you can visualise the evidence across the therapeutic areas in a tree format that represents the classification of diseases by subtypes. Therapeutic areas have a square symbol (e.g. Genetic disorders), while the diseases (e.g. multiple sclerosis) are represented as circles. The squares and circles are colour coded in blue, and the darker the blue, the stronger the association:



For all these three different views, you have the option to filter the data according to Data type or Therapeutic area. You can, for example, explore all diseases of the Reproductive system associated with AR. In the June 2019 release of the Open Targets Platform, there are 113 diseases of the Reproductive system disease, such as 'polycystic ovary syndrome' and 'Rare genetic female infertility'.

You can view the same information in a bubbles view. Click on any of the smaller bubbles, which represent specific disease. This will bring a pop-up window with the flower plot for an overview on the evidence used for the association) and the following options:



- "Profile" to get to the profile page of the disease
- "Associations" to get all the target associations for the disease
- "View evidence details" to see the underlying data for the associations

Note: in the associations page for AR, you can click on "View AR profile" to see the annotations for this target, such as Target safety and Target tractability. These can help you to prioritise which targets to pursue for downstream analysis.

HANDS-ON EXERCISES

Open Targets Platform

Exercise 1: Durvalumab and non-small cell lung carcinoma

BACKGROUND
AstraZeneca and MedImmune have recently announced the final overall survival results for the Phase III MYSTIC trial, a randomised, open-label, global trial of durvalumab, a human monoclonal antibody that binds to PD-L1 and blocks the interaction of PD-L1 with PD-1 and CD80, countering the tumour's immune-evading tactics and releasing the inhibition of immune responses.

QUESTIONS
a) How many targets have associations involving this drug in the Open Targets Platform? Note: the match between the search term (durvalumab) and one of the targets is partial; the drug name matches the protein LUMA. In this case, the drug is not in clinical trials for the modulation of this protein.

b) Let's now focus on the target returned in the result of the previous question. Which data types and sources support the association of this target with non-small cell lung carcinoma?

c) Looking closely at the drug evidence for the association between this target and non-small cell lung carcinoma, click on the 'Drugs' cell in the table in the associations page. You are now in the evidence page and the 'Drugs' section is opened. How many clinical trials using 'durvalumab' have been withdrawn by the FDA?

d) Now click on the hyperlinked drug name. You will be redirected to the 'drug summary page'. When was this drug first approved? What is its trade name? In addition to non-small cell lung carcinoma, can you name other diseases where this drug is in clinical trials for? Have any adverse effects been reported for this drug?

e) Let's go back to the evidence page. Are there any papers from 2019 that support the link between this target and non-small cell lung cancer through **text mining**? Can you download this information?

Exercise 2: Advancing research in the field of Crohn's disease

BACKGROUND
Crohn's disease is a lifelong condition in which parts of the digestive system become inflamed. While the cause of Crohn's disease is unknown, it is believed to be due to a combination of environmental, immune and bacterial factors in genetically susceptible individuals. Therapeutic drugs are available to treat some of its symptoms but no cure has yet be found.

Use the Open Targets Platform to answer the following:

QUESTIONS
a) How many targets are associated with Crohn's disease? How many of those are involved in the interleukin-4 and 13 signaling pathway (immune system) and are associated with the disease based on GWAS catalog evidence only?

What is the only target from the list above that is classified as a membrane receptor?

b) Look at the evidence behind the association between this membrane receptor and the disease. Can you find the actual variant implicated in this association? Explore the Browser tab to see this information in the context of the human genome.

c) Is there any evidence showing a change in the RNA expression of this target in colonic mucosa? Does the change correspond to a decrease or increase in RNA levels when comparing tissues from patients diagnosed with Crohn's disease vs normal individuals? Note: an increase in RNA expression in the disease tissue versus normal would suggest the gene is up-regulated in the disease state, whereas a decrease would suggest the opposite i.e. the gene would be down-regulated in the diseased tissue.

Open Targets Genetics

Exercise 3: Demonstrating the role of genetics supporting an existing drug

BACKGROUND
Pursuing drug targets that have supporting genetic evidence could double the clinical development success rate (Nelson et al. 2015). Genetically-supported targets could be used to choose which indications and targets to focus on. In this exercise, we use Open Targets Genetics to demonstrate retrospectively the role of genetic variants in HMGCR for the success of modulating this target with an approved drug. Search for LDL cholesterol and pick Willer CJ (2013) Nat Genet.

QUESTIONS
a) How many loci are independently associated with LDL cholesterol at genome-wide significance (p-value < 5e-8) in this study?

b) How many independent associations can you find in chromosome 5?

c) Which genes are functionally implicated by 5_75329662_C_A

d) Can you rank these genes by the distance to the canonical TSS (transcript start site), from closer to farther away from the TSS? *Note: by default, the ranking is set for the overall V2G score, from the highest to the lowest*.

e) What functional evidence supports the link of this variant to gene *GCNT4*?

f) Click on the eQTL tab in the same table to view tissue and direction of effect.

- In which tissue is there eQTL evidence for HMGCR?
- What is the direction of effect?

g) 'LDL cholesterol' (Study ID GCST002222) is the most significantly associated trait with this variant as shown in the PheWAS-like plot and table. What other traits are associated with this variant at

phenome-wide significance? *Note the direction of effect by the triangles pointing upward or downward in the PheWAS-like plot (and corresponding negative or positive beta coefficient in the table).*

h) In the table below the PheWAS plot, search for the UK Biobank study 'high cholesterol |Non-cancer illness code, self-reported'. Scroll along the table and click on the 'Locus Plot' button

In the locus plot, use the drop-down to toggle from the default selection, 'Show expansion by LD and fine-mapping' LD and fine mapping', to 'Show expansion by fine mapping only'.

The table below the locus plot displays the variants tagging this lead variant (5_75329662_C_A, rs7703051) e.g. 5_75330257_T_C and 5_75359901_C_T, and the genes functionally implicated by these tag variants, e.g. HMGCR. Let's click on the gene name to go to the gene page in Open Targets Genetics. In addition to 'LDL cholesterol (Willer CJ 2013), which other studies (or traits) are associated with HMGCR?

i) From Open Targets Genetics, can you find the links to Open Targets Platform, and learn if there are already approved clinical trials (Phase IV) where this target is modulated by a drug (or drugs)?

j) Can you name the drugs in Phase III, status completed, where HMGCR is modulated to treat myocardial ischemia?

Exercise 4: Identifying diseases and molecular QTLs that colocalise with a disease-associated signal

BACKGROUND
VEDOLIZUMAB is an approved (Phase IV) drug targeting ITGA4 to treat Crohn's disease. We will use Open Targets Genetics to find if independent association signals at ITGA4 are consistent with having a shared casual variant. This is known as colocalization.

QUESTIONS
a) Which studies have evidence of colocalisation with molecular QTLs for ITGA4 (*hint: search for this gene and look for the colocalization studies table*)?

b) One of the lead variants for study GCST004132 (Crohn's disease) is is 2_181443625_A_G. Click on any of the **Colocalisation** buttons in the table. *Note all buttons will lead to the same colocalisation page between study GCST004132 and the variant (i.e. 2_181443625_A_G)*.

Let's first look at the 'QTL colocalisation' section. What molecular traits colocalise with this Crohn's disease (de Lange KM, 2017) signal? What tissues is the colocalisation reported?

d) Let's now look at the 'GWAS study colocalisation'. What GWAS traits colocalise with Crohn's disease (de Lange KM, 2017) at this locus (GWAS traits are labelled followed the suffix GCST)? *Note the different posterior probability hypothesis tested, H3 (independent traits) and H4 (share associations between traits) and the ratio between the two.*

e) Which shared variants are most likely causal between two colocalised studies e.g. Crohn's disease (de Lange KM, 2017) and inflammatory bowel disease (de Lange KM, 2017)?

f) Let's now compare the 'Crohn's disease' signal with and the QTL 'CEDAR: ITG4A in Monocyte CD14'. Do they look signals look the same in the regional plots?

f) Can you identify the set of likely causal variants at this locus based on the Crohn's disease and ITG4A Monocyte signals? Download this information in "TSV" for further prioritization analysis.

i) Which studies have evidence of colocalisation with molecular QTLs for HMGCR?

QUICK GUIDE TO DATABASES

Here is a list of databases and projects that may be useful for you. Some of them are used as data sources for gene-disease associations available through our Open Targets Platform.

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/

GERMLINE 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/

Gene2Phenotype - The data in Gene2Phenotype (G2P) provides evidence of genetic variants that are manually curated from the literature by consultant clinical geneticists in the UK. This is provided by DECIPHER, a database of genomic variants and phenotypes in patients with developmental disorders.
https://www.ebi.ac.uk/gene2phenotype

Genomics England PanelApp - The Genomics England PanelApp is a knowledgebase that combines crowdsourcing of expertise with curation to provide gene-disease relationships to aid the clinical interpretation of genomes within the 100,000 Genomes Project.
https://panelapp.extge.co.uk/crowdsourcing/PanelApp/

PheWAS Catalog
The PheWAS (phenome-wide association studies) resources provide associations between a genetic variant and multiple phenotypes. It contains clinical phenotypes derived from the electronic medical record (EMR)-linked DNA biobank BioVU by the Center for Precision Medicine at the Vanderbilt University Medical Center.
https://phewascatalog.org/

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/

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 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/

COSMIC is also the database that provides us with the cancer hallmarks:
https://cosmic-blog.sanger.ac.uk/hallmarks-cancer/

IntOgen - It provides evidence of somatic mutations (driver mutations), genes and pathways involved in cancer biology from 6,792 samples across 28 cancer types.
https://www.intogen.org/search

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

PATHWAYS & SYSTEMS BIOLOGY

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/

SLAPenrich – It's a statistical framework for the identification of significantly mutated pathways, at the sample population level. We include in the Open Targets Platform the data obtained using SLAPenrich on somatic mutations from the The Cancer Genome Atlas across 25 different cancer types and a collection of pathway gene sets from Reactome.
https://saezlab.github.io/SLAPenrich/

PROGENy – PROGENy (Pathway RespOnsive GENes) is a linear regression model that calculates pathway activity based on consensus gene signatures obtained from perturbation experiments. We use PROGENy (PMID:29295995) for the systematic comparison of pathway activities between normal and primary samples from The Cancer Genome Atlas (TCGA). We include in our Open Targets Platform sample-level pathway activities inferred from RNA-seq for 9,250 tumour and 741 normal TCGA samples from 14 tumour types, and compute differential pathway activities between matched normal and tumour samples. We cover the following pathways: EGFR, hypoxia, JAK.STAT, MAPK, NFkB, PI3K, TGFb, TNFa, Trail, VEGF, and p53. See PMID:29295995 for more details.

Sysbio – This data source includes six gene lists curated from four systems biology analysis papers. These publications integrate different types of data to identify key drivers (or regulators) in the following diseases (or phenotypes):

- Inflammatory bowel disease (PMID:28892060)
- Coronary heart disease (PMID:23539213)
- Late-onset Alzheimer's disease (PMID:23622250)
- Cognitive decline of Alzheimer's disease (PMID:29802388)

In the Platform, we have 406 evidence strings used to associate targets to any of the diseases or phenotype above.

CRISPR – This data is a subset of the gene essentiality data generated in cancer cell lines (PMID:30971826). The top genes prioritised after CRISPR screens were combined with target tractability and genomic marker information to give rise to a list of 624 genes from 19 different cancer types and 1,846 target-disease associations.

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/

ANIMAL MODELS

Phenodigm - Phenodigm is an algorithm developed by Damian Smedley at the Wellcome Sanger Institute that use a semantic approach to map between clinical features observed in humans and mouse phenotype annotations. The results are made available on the IMPC portal:
https://www.mousephenotype.org