**Open Targets: integrating genetics, omics and chemistry for drug discovery**

Cancer Research UK

AstraZeneca Antibody Alliance



14th January 2020

Denise Carvalho-Silva

Open Targets

EMBL-EBI

United Kingdom

**Welcome to our training session!**

This booklet is based on [release 3.15](https://docs.targetvalidation.org/release-notes#release-3-15-2019-11-28) (2019-11-28) 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

Questions or suggestions?

[support@targetvalidation.org](mailto:support@targetvalidation.org)

**TABLE OF CONTENTS**

OVERVIEW 4

INTRODUCTION TO OPEN TARGETS 5

OPEN TARGETS PLATFORM: WALKTHROUGH 8

Demo 1: Disease centric workflow 8

Demo 2: Target and disease annotations 16

Demo 3: Target centric workflow 24

HANDS-ON EXERCISES 28

Exercise 1: Durvalumab and non-small cell lung carcinoma 28

Exercise 2: Advancing research in the field of pancreatic carcinoma 29

EXTRA HANDS-ON EXERCISES 30

*Exercise E1: The EGFR gene, a receptor tyrosine kinase* 30

*Exercise E2: Filtering prostate cancer associations based on a list of eight targets* 31

QUICK GUIDE TO DATABASES 33

# 

# 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 and the partnership was rebranded to Open Targets in April 2016. Since then we have welcomed three new partners: Takeda in 2017, and Celgene and Sanofi 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. However, 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
* The Open Targets Platform
* How to browse the web interface of the Platform
* 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

Our core bioinformatics, data pipelines and core Genetics team have developed a suite informatics resources, namely Open Targets Platform, Open Targets Genetics, and LINK (the Open Targets LIterature coNcept Knowledgebase).

More details on our projects can be found on the [Scientific Overview](https://www.opentargets.org/science/) page and the complete list of publications and informatics resources is available on the [Resources](https://www.opentargets.org/resources/) page. Two publications that are worth noting specially researchers and drug discovery scientists interested in cancer are the following:

* [Prioritisation of oncology therapeutic targets using CRISPR-Cas9 screening](https://www.nature.com/articles/s41586-019-1103-9)
* [Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets](https://www.nature.com/articles/s41467-019-13805-y)

**Informatics resources**

The first Open Targets informatics resource (the Open Targets Platform) to provide easy access to data relevant to drug target identification and selection by a diverse audience of users was launched in December 2015. The Open Targets Platform is a web application that integrates 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](https://docs.targetvalidation.org/data-sources/data-sources) to associate genes and diseases. Similar data **sources** are combined into the following data **types** in the Open Targets Platform:

Genetic associations

Somatic mutations

Drugs

Pathways & systems biology

RNA expression

Text mining

Animal models.

Different types of evidence (e.g. SNPs, scientific literature) from the data sources are used to compute the [association score](https://docs.targetvalidation.org/getting-started/scoring), which depends on the frequency of evidence, the confidence on the evidence and on its severity (for example, if a genetic variant changes the amino acid sequence of a protein, its severity is higher than variants that lie in the intronic region of a gene). We then aggregate the evidence score using the sum of the [harmonic progression](https://en.wikipedia.org/wiki/Harmonic_progression_(mathematics)) to obtain the score at the data source and data type levels, as well as the overall score. The association score can be used to rank target and disease associations in the Platform.

The latest release of the Platform (November 2019) contains:

* 27,069 targets
* 13,579 diseases
* 6,336,307 target-disease associations
* 8, 919,232 pieces evidence to associate a target with a disease

The Open Targets Platform is an open source and open access easy-to-use tool that can be applied in a variety of use cases in academia and pharma industry.

**What can you do with the Open Targets Platform?**

* Find targets associated with any disease
* Find diseases associated with any given target
* Find the evidence supporting target-disease associations
* Find annotations for targets e.g. tractability data, safety data
* Find annotations for diseases e.g. all drugs in trials or marketed
* Use the batch search tool to search for a list of up to 200 targets
* Carry out more complex queries using the [REST API](https://docs.targetvalidation.org/programmatic-access/rest-api)
* Download all evidence and association objects for all human genes and diseases to carry out your own downstream analysis

**Connect with us**

* [Open Targets Blog](file:///C:\Users\gk680303\Downloads\blog.opentargets.org)
* Follow us on [Twitter](http://www.twitter.com/targetvalidate), [Facebook](https://www.facebook.com/OpenTargets/), [LinkedIn](https://www.linkedin.com/company/centre-for-therapeutic-target-validation)
* Watch our video tutorials and short animations on [YouTube](https://www.youtube.com/channel/UCLMrondxbT0DIGx5nGOSYOQ)

# OPEN TARGETS PLATFORM: WALKTHROUGH

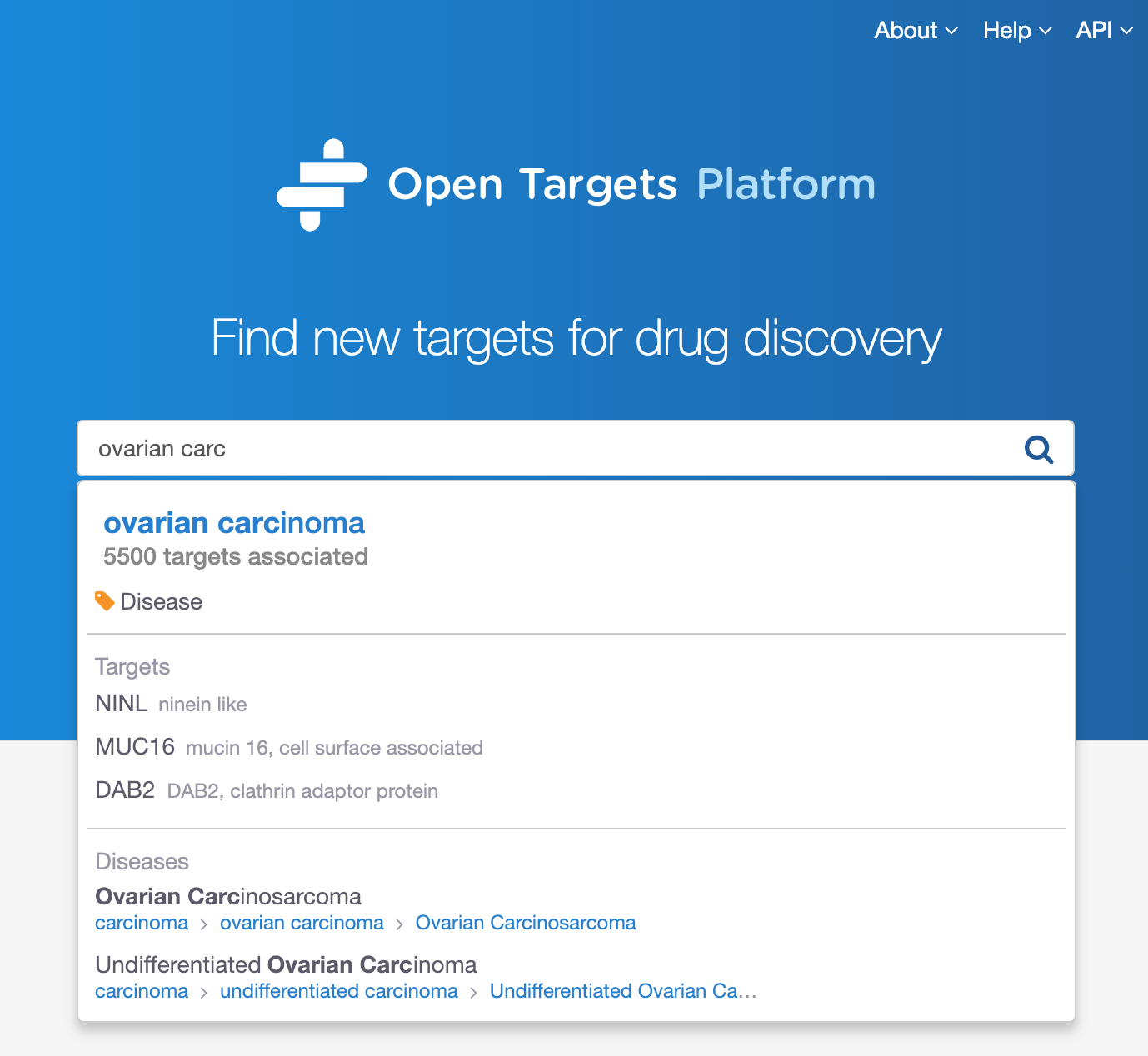
We will guide you through the website using **ovarian carcinoma**, then we will explore the evidence associating ERBB2with this disease.

The following points will be addressed during the walkthrough:

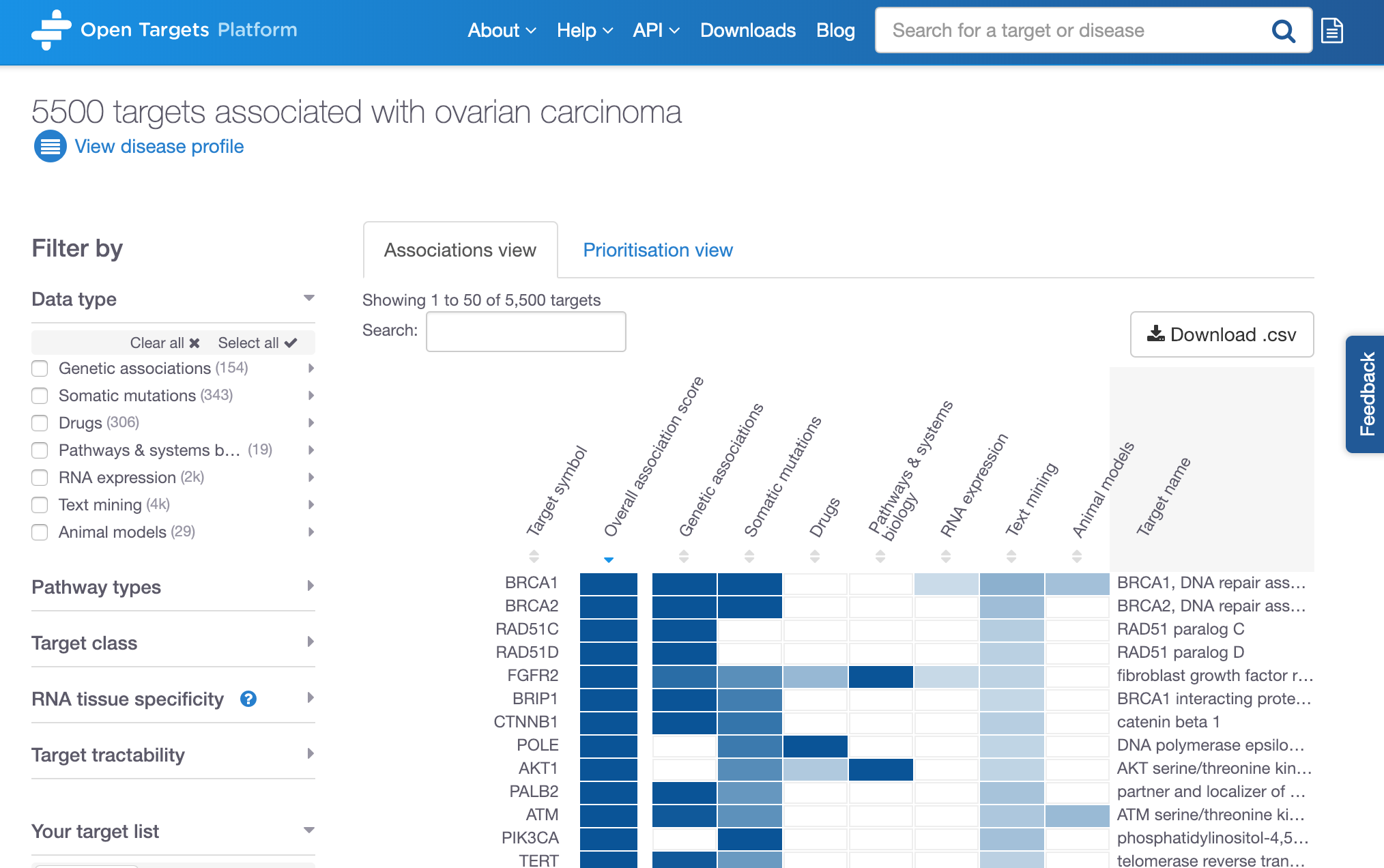
* + How to find targets associated with ovarian carcinoma
  + How to filter down the number of targets based on specific types of evidence, pathway types, target class, etc
  + How to find out how strong the association between ERBB2 and ovarian carcinoma is
  + How to view the evidence that supports 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: Disease centric workflow****

Go to [www.targetvalidation.org](http://www.targetvalidation.org):



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



Data types

(Somatic mutations,

Drugs, etc)

Total number of targets associated with ovarian carcinoma

Click on the arrows to sort the rows by (increased or decreased) score values of individual data types.

Filter the results

The current release of the Open Targets Platform (November 2019) lists 6432 targets associated with ovarian carcinoma.

The table in the associations page (see the image above) shows the best targets at the top of the table by default. The first target i.e. BRCA1 is the target that contains the most compelling set of supporting evidence behind the association with ovarian carcinoma. This is summarised by the overall association score. The association score can vary from 0 to 1; the closer to 1 the more evidence we have for an association. This score is computed for each piece of evidence that is used to support the association. Then individual **evidence scores** within a data source are aggregated into one **data source score** followed by the aggregation of data source scores into a common **data type score**. The Overall association score column in the above table is the aggregation of all data source scores.

*Note 1: whenever we aggregate scores, we apply a harmonic sum.*

*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 capped at 1 for all those three targets, we have more evidence available for BRCA1; hence it comes before BRCA2 in the ranked table.*

We also give different weight to different data types whilst computing the association score. RNA expression, animal models and text mining are all down weighted by a factor of 0.2, whereas Sysbio, PROGENy and SLAPenrich are down weighted by a factor of 0.5. Check the help documentation for more details on the [association score](https://docs.targetvalidation.org/getting-started/scoring) and [data sources](https://docs.targetvalidation.org/data-sources/data-sources) available in the Open Targets Platform.

Note that you can sort the table by alphabetical order of target names, or by the association score values, either overall score or per data type e.g. Genetic associations, Drugs, Text mining, etc.

The association table listing all > 6,400 targets associated with ovarian carcinoma 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 below:

**1) Data types:** we collect evidence from various data 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

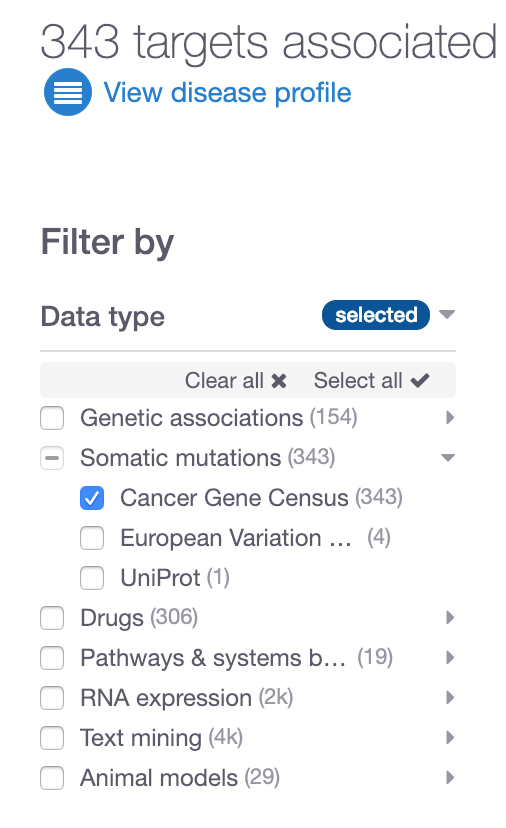
Pathways & systems biology

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

…

Note that some of these pathways can be further subdivided into different subtypes, e.g. immune system can be broken down into neutrophil degranulation, Antigen processing and many others:

A screenshot of a cell phone

Description automatically generated

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

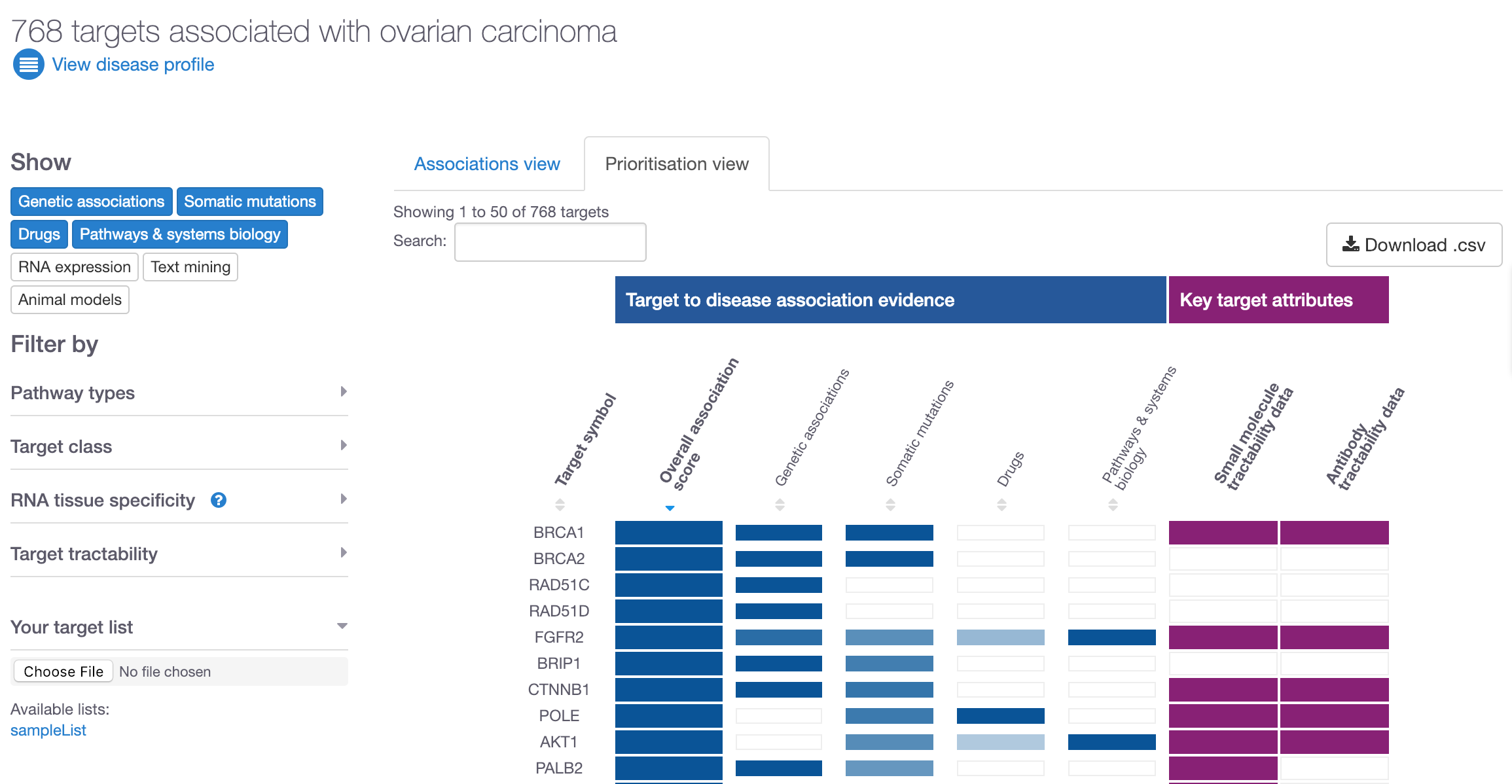
**4) RNA tissue specificity:** RNA tissue specificity: the tissue specificity of a target is computed as the number of standard deviations from the mean of the log RNA expression of the target across the available tissues. This is a standard z-score calculation. A target is considered to be tissue specific 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: t**ractability of a target is the confidence that we can identify a modulator that interacts with the target to elicit a desired biological effect. Currently we assess whether the targets can be modulated by a Small molecule or Antibody.

**6) 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 that in this associations page, there are two views, the Associations view and the Prioritisation view. The latter shows whether the available targets can be modulated by either small molecule, antibody, both modalities or neither:



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 from 6432 to 695. Let’s now focus on ERBB2. Search for this gene symbol using the search box above the top of the table:

A screenshot of a cell phone

Description automatically generated

Click on any of these coloured cells to go to the evidence page for ERBB2 in ovarian carcinoma.

Search for ERBB2

Cancer Gene Census

selected

Click on any of the blue cells in the table to go to the evidence page:

A screenshot of a social media post

Description automatically generated

In the evidence page, you can explore the evidence used for the association between ERBB2 and ovarian carcinoma.

The coloured petals on the flower plot represent the data types that support this association namely:

Genetic associations

Somatic mutations

Drugs

Pathways

Text mining

**Note**: If you wish to suggest data types or resources we could/should incorporate into our Platform, please get in touch:

support@targetvalidation.org

Grey areas in the flower plot above indicate there is no evidence for those data types.

Let’s now scroll down on the page and expand the tabs available, for example, Somatic mutations (with links to COSMIC, which provides us with Cancer Gene Census data) and Drugs, to find if there are drugs currently in clinical trials targeting ERBB2in patients with ovarian carcinoma (or children terms of this disease). There are currently four drugs currently listed in different phases of different clinical trials. Do you know why there are seven entries, although only four unique drugs? Are there any drugs in late phases of clinical trials i.e. phase III?

Have a look at the text mining results to find papers used as evidence for ERRB2-ovarian carcinoma associations.

Once we explore the evidence, we can move on to look at target and disease annotations.

## ****Demo 2: Target and disease annotations****

Let’s now scroll back up to the top of this page and click on the “ERBB2” link:

A screenshot of a social media post

Description automatically generated

Click on the target link to get to the target profile page

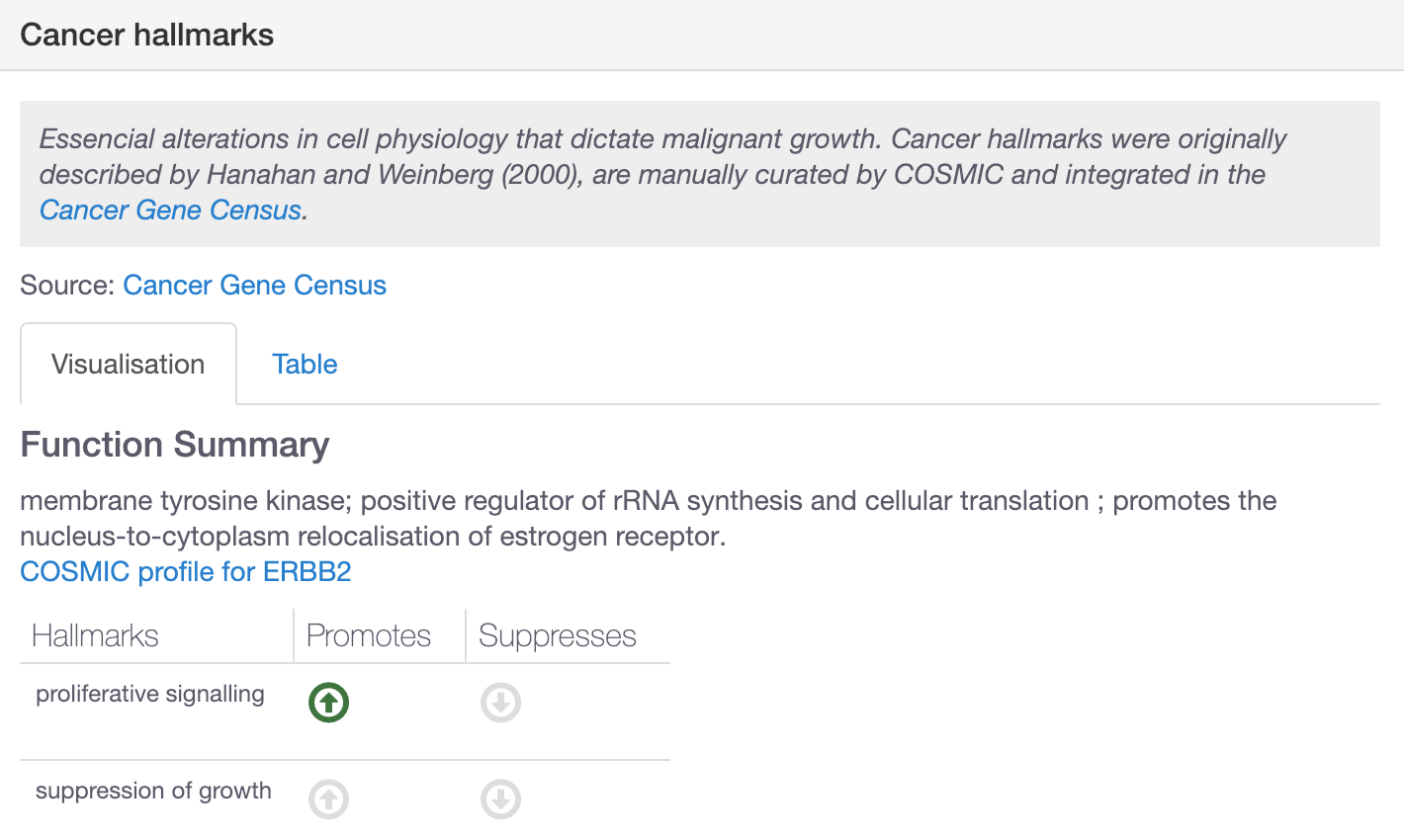
Click on ERRB2 in the above image to be redirected to the target profile page of this gene:

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

A screenshot of a social media post

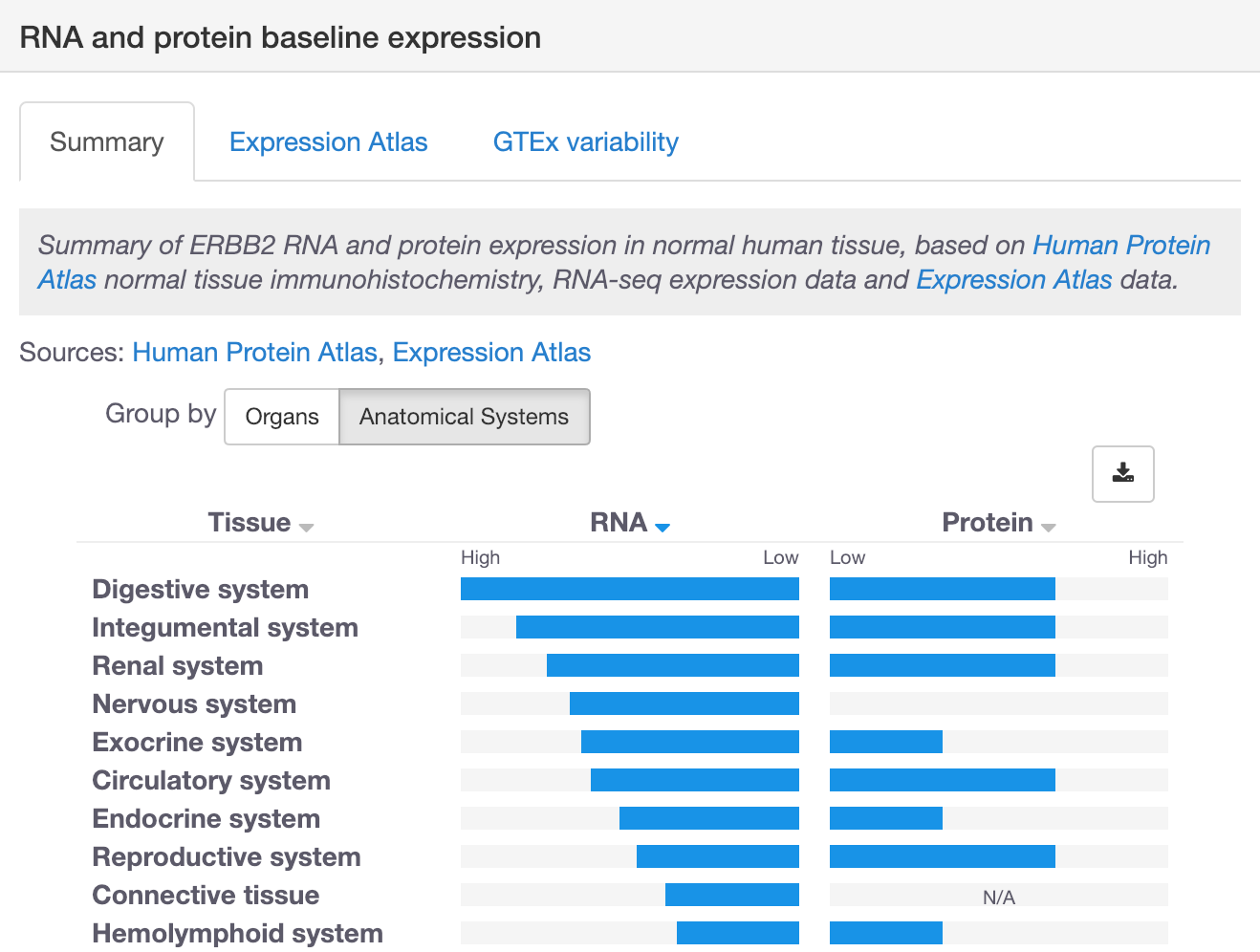
Description automatically generated

This is the target profile page where you can find gene specific information for ERRB2, e.g. RNA and protein baseline expression levels, protein structure, gene ontology terms, information on tractability of ERRB2, cancer biomarkers, cancer hallmarks, and much more.



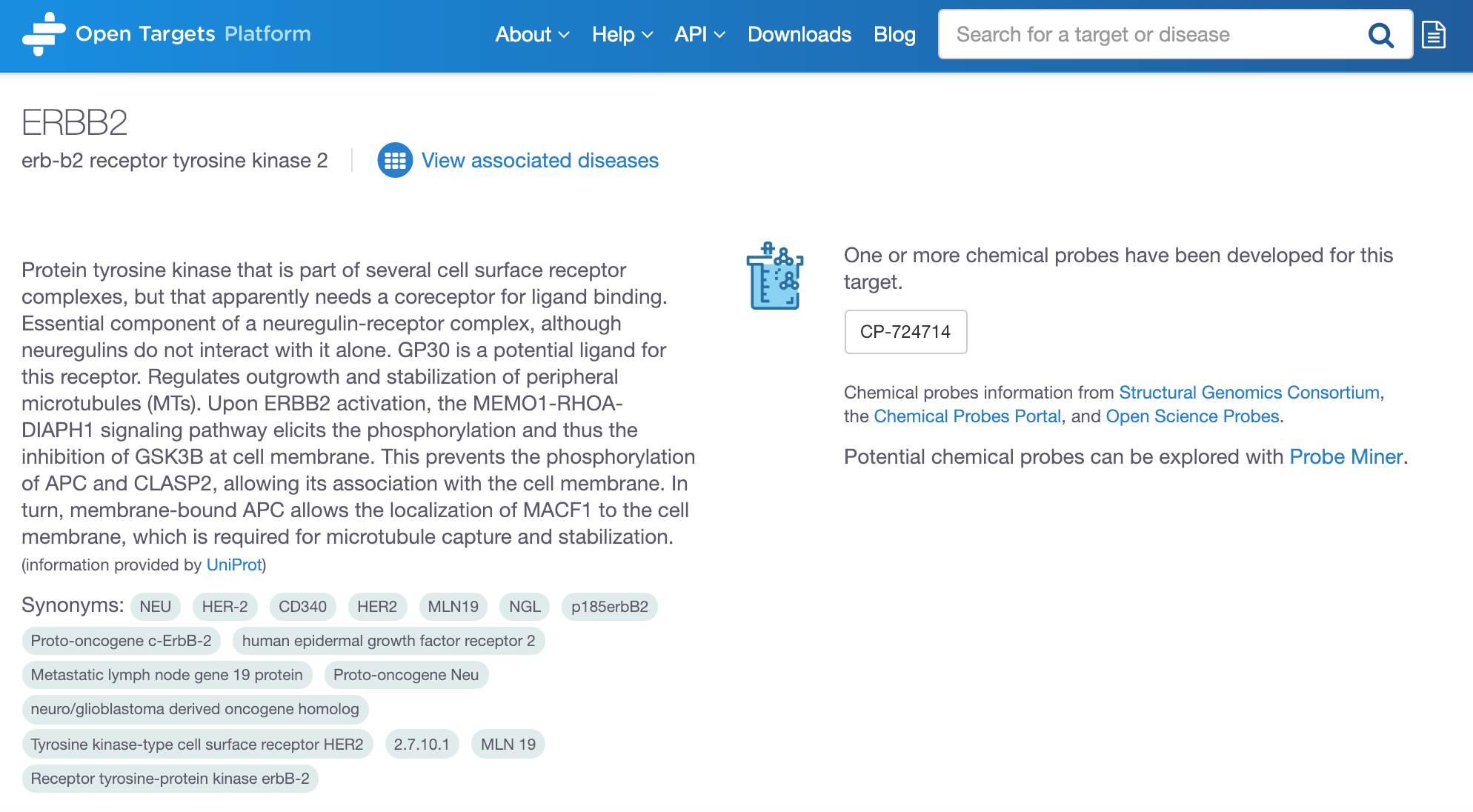
Let’s now expand the RNA and protein baseline expression to find out in which organs or anatomical systems ERRB2 is highly expressed.

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

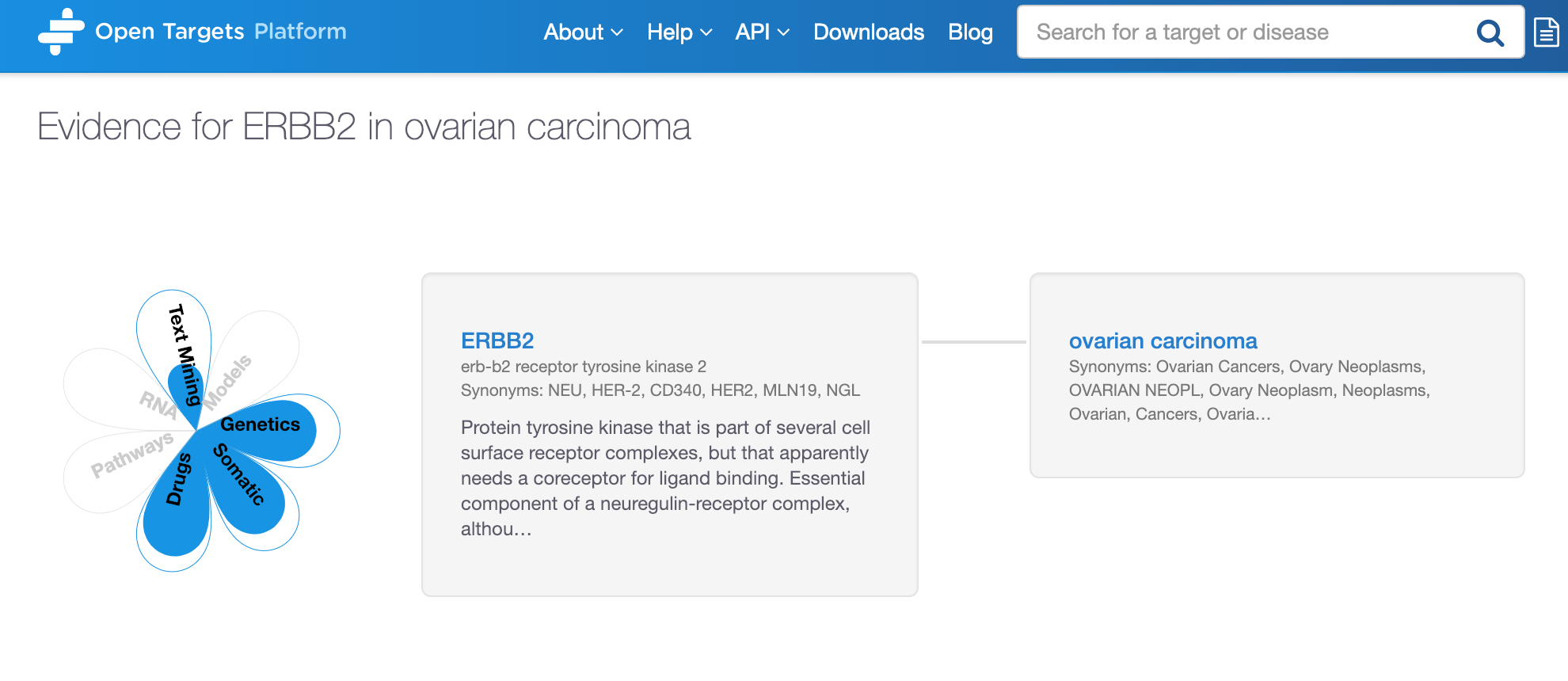


In the Summary tab, you can compare the mRNA and protein expression side by side and learn that the expression in the reproductive system is higher at the protein than RNA level. You can click on the tissue names to get further granularity such as different different tissues in the digestive system such as Esophagus, small intestine, etc:

Note that from the target profile page you can also explore other diseases associated with ERBB2, apart from ovarian carcinoma, if you click on View associated diseases:



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



Click on the disease link to get to the disease profile page

Click on the disease name and explore the annotations for ovarian carcinoma:

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

Scroll down to view the disease ontology (disease concepts and relationships) under the Classification tab. This comes from the EFO (Experimental Factor Ontology), an ontology developed and maintained by EMBL-EBI.

A screenshot of a social media post

Description automatically generated

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 any disease:

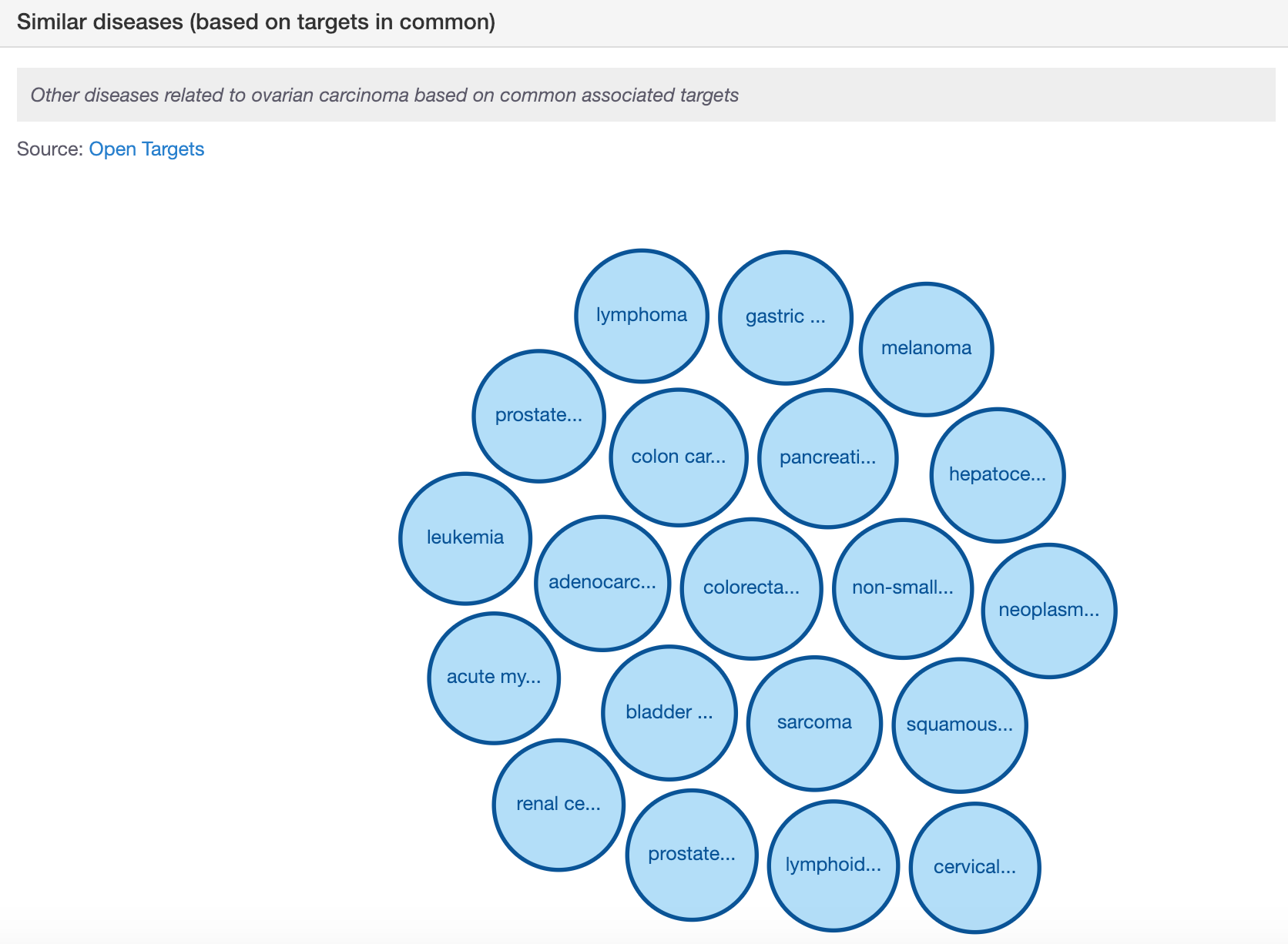
A screenshot of a cell phone

Description automatically generated

By taking into account the disease ontology and its subtype relationships, 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 [direct versus indirect](https://blog.opentargets.org/direct-versus-indirect-evidence-should-you-care/) blog post.

Let’s now check the Similar diseases (based on targets in common) tab in the disease profile page:



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.

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.

You can 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):

A screenshot of a cell phone

Description automatically generated

## ****Demo 3: Target centric workflow****

You can also use the Open Targets Platform from a target standpoint. Let’s search for AR and explore the visualisations and some the functionalities available in the Open Targets Platform.

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.*

* 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.

* 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 or breast disease” associated with this target.

In the latest release of the Open Targets Platform, there are 197 of them. These are some examples: ‘cervical adenocarcinoma’ and ‘prostate adenocarcinoma’.

You can view the same information in a bubble view. Click on any of the smaller bubbles, which represent a specific disease. This will bring a pop-up window with the flower plot (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 evidence for the associations

On the other hand, if you are in the tree view, click on any of the square or circles for a similar pop-up menu.

---------------------------------------------------------------------------------------------------------

End of the walkthrough

---------------------------------------------------------------------------------------------------------

# HANDS-ON EXERCISES

## 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 randomized trial of durvalumab. Durvalumab is a human monoclonal antibody that binds to CD274 and blocks its interaction with CD279, countering the tumour's immune-evading tactics and releasing the inhibition of immune responses.

QUESTIONS

a) You can search for durvalumab (or any other drug) using the Open Targets Platform. How many targets are returned in the search results when searching for this drug name? Note the extent of the match between the search term and the returned targets in the search results page. For one of the targets, the match is a false negative (the word LUMA gets matched to durva**LUMA**b).

b) Let’s now focus on target CD274 and check the diseases associated with it. Which data sources support the association of this target with non-small cell lung carcinoma? In addition to durvalumab, are there other drugs currently in **Phase III** of clinical trials, whose **status** is “recruiting”? Can you name this other drug? Is it an antibody, small molecule or other drug modality?

c) Is there any RNA expression evidence for the association between non-small cell lung carcinoma and CD274 that shows a decreased activity when comparing primary tumour with adjacent normal tissue in the lung?

d) Go back to the Drug section and click on the durvalumab link in the Drug information table. You are now in the Drug summary page. When was this drug first approved? What is the mechanism of action of this drug? Can you name some of the most significant adverse events linked to this drug based on the FDA pharmacovigilance data? Are there other diseases where this drug is associated with? Can you find the list papers from the scientific literature on this drug?

## Exercise 2: Advancing research in the field of pancreatic carcinoma

BACKGROUND

Pancreatic acinar cell carcinoma (PACC) is a rare neoplasm and has been considered a cancer with poor prognosis. Patients who present with PACCs usually have non-specific symptoms. Their complaints consist of abdominal pain or discomfort, nausea, vomiting, weight loss, and diarrhea.

QUESTIONS

a) How many targets that are membrane receptors (therefore target class amenable to antibody drug discovery) are associated with pancreatic acinar cell carcinoma based on somatic mutations (data type) only? Are these targets tractable by an antibody? What is the information supporting this assumption/prediction?

b) Let’s now have a look at another target associated with PACC with strong support from the evidence available in the Open Targets Platform. The top associated target is SMAD4. Let’s focus on this target and explore its profile page to answer the following:

- Which amino acids correspond to an antibody binding sequence (i.e. antigenic sequence) in the protein encoded by this gene?

- Can you name some of the pathways this protein is involved at?

- Which tissue has the highest level of protein baseline expression? Does this correspond to the highest expression at the RNA level as well?

c) Let’s now explore the disease profile page for PACC. Can you name some of the diseases that are similar to PACC based on targets in common? Which targets are in common between PACC and colon adenoma? What is the evidence for the associations between SMAD4 and these two similar diseases (i.e. PACC and colon adenoma)?

# EXTRA HANDS-ON EXERCISES

## *Exercise E1: 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)* *Are there any target safety concerns when modulating this target? Can you find the publications citing these concerns?*

*b) Can you list a few examples of molecular functions (according to the Gene ontology) for the EGFR protein?*

*c) What are some of the phenotypes associated with this gene in mice? Is there any neoplasm associated with Egfr in mice?*

*d) Is EGFR involved in the promotion of proliferative signalling. This is one of the well-known cancer hallmarks manually curated by COSMIC and available via the Cancer Gene Census.*

*e) The Open Targets Platform has integrated cancer biomarkers of drug responses and their levels of clinical significance. These can be mutations in the EGFR protein that can confer drug resistance in certain types of cancers. Can you name some biomarkers that seem to be associated with cetuximab resistance? Can these biomarkers be used in different types of cancers?*

## *Exercise* *****E2:***** *Filtering prostate cancer associations based on a list of eight targets*

BACKGROUND

A drug discovery scientist at CRUK Therapeutic Discovery Laboratories in London has a list of eight targets that seem to be associated with prostate carcinoma based on literature reviews. These are CHEK2, AR, CDKN1A, CDK12, PTEN, TSC2, KLF6, and MXI1. Copy and paste these targets, one per line, and save this in plain text file (.txt).

Go to the Open Targets Platform and search for prostate carcinoma. In the association page listing the targets for this disease, scroll down the page till you see the filter option Your target list. Click on the Choose file button, search for your file (saved above) and upload this to the Open Targets Platform. Once you have done this, you will be able to answer the following:

QUESTIONS

a) Which of those eight targets have higher levels of mRNA expression in the prostate gland than in any other tissue? Note: this is known in the Open Targets Platform as RNA tissue specificity. Prostate gland is under the group Anatomical Systems.

Let’s now explore the evidence for the association between one the targets and prostate carcinoma. The target is the androgen receptor (AR).

b) Are there any variants or mutations that cause a missense change in the AR protein and that has clinical significance of being pathogenic? Can you find the publications that support this?

c) Are there any drugs in Phase 0 of clinical trials targeting AR in patients suffering from prostate carcinoma? Can you retrieve the clinical trials for this data?

d) Are there any drugs in Phase 0 of clinical trials targeting AR in patients suffering from prostate carcinoma? Can you retrieve the clinical trials for this data?

*e) Can you name a few publications used as Text mining evidence for the AR-prostate carcinoma association where the target and disease names co-occur in the same sentence in the title of the paper?*

# 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](https://docs.targetvalidation.org/data-sources/data-sources) for gene-disease associations available through our Open Targets Platform.

**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 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/](http://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>

**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/>

**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](https://saezlab.github.io/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 ([Schubert et al](https://www.nature.com/articles/s41467-017-02391-6.epdf?author_access_token=16QkzhJ3OA3qJDqBw_GvGdRgN0jAjWel9jnR3ZoTv0NBFLUVI-ebH2AmtFlR1ykSPIho7ETJXL7VqZFC4zGtU0BaeoZncGrwx3ZW24lfVqvbSWqsQKaUXFTi_c-4pgcpX-1qerWYlkG6sha8rhrnMg%3D%3D)) 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 [Schubert et al (2018)](https://europepmc.org/abstract/MED/29295995) for more details.

**SYSBIO -** Sysbio 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](https://europepmc.org/abstract/MED/28892060))
* Coronary heart disease ([PMID:23539213](https://europepmc.org/abstract/MED/23539213))
* Late-onset Alzheimer's disease ([PMID:23622250](https://europepmc.org/abstract/MED/23622250))
* Cognitive decline of Alzheimer's disease ([PMID:29802388](https://europepmc.org/abstract/MED/29802388))

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

**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 Trust 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>