Open Targets: integrating genetics and genomics for drug discovery

C4X Discovery

18th June 2019

Denise Carvalho-Silva, PhD

EMBL-EBI | Open Targets
United Kingdom

Notes

This booklet is based on the April 2019 release (19.04) of the Open Targets Platform. These are some useful links:

1) Open Targets - Scientific overview
https://www.opentargets.org/science/

2) Open Targets Platform help
https://docs.targetvalidation.org/

3) Open Targets Platform FAQs
https://docs.targetvalidation.org/faq/frequently-asked-questions

4) Videos and animations
https://tinyurl.com/opentargets-youtube

5) Read our latest publication
http://bit.ly/cite-us

Questions or suggestions?
support@targetvalidation.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
- The Open Targets Platform and Open Targets Genetics
- How to browse Open Targets web resources
- Overview of the programmatic access of 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 and data pipelines team has developed the Open Targets Platform to provide easy access to data relevant to drug target identification and selection by a diverse audience of users. More recently, we have also launched Open Targets Genetics, for the exploration variant-gene-trait associations from UK Biobank and GWAS Catalog.

More details on our projects can be found on our Scientific Overview page. The first set of results from our experimental work on the Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens is now published.

The Open Targets Platform

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 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 an association score, which depends on the frequency of evidence, the confidence and severity (e.g. does the SNP change the amino acid of the target protein?). We then aggregate the evidence score using the sum of the harmonic progression 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 (April 2019) contains:

- 28,501 targets

- 10,419 diseases

- 3,303,824 associations between targets and diseases

- 7,209, 475 evidence

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.

What can you do with the Open Targets Platform?

- Search for a disease and find its associated targets based on genetics, transcriptomics, drug information, text mining, etc
- Search for a target and find its associated diseases based on biological and chemical evidence
- Delve deep into the evidence supporting target-disease 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](#)

<p align="center">**OPEN TARGETS PLATFORM: WALKTHROUGH**</p>

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:

- o How to find targets associated with ovarian carcinoma
- o How to filter down the number of targets based on specific types of evidence, pathway types, target class, etc
- o How to find out how strong the association between ERBB2 and ovarian carcinoma is
- o How to view the evidence that supports this association
- o How to find other diseases associated with ERBB2
- o How to visualise the ERBB2 gene and its genetic variants on the human genome
- o How to find drugs currently in clinical trials for ERBB2
- o How to filter the Open Targets 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 (April 2019) lists 5614 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) for each piece of evidence that is used to support an association (evidence score) e.g. a single SNP
2) for all pieces of evidence within a data source (data source score) e.g. all SNPs from GWAS Catalog
3) for all data source scores within a 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 page to find out more about our data sources: https://docs.targetvalidation.org/data-sources/data-sources

The association table listing all 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 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. Currently we assess whether the targets can be modulated by a Small molecule or Antibod

Tractability prediction for small molecules is further subdivided into:

- Discovery precedence

- Predicted tractable

- Clinical precedence

Tractability prediction for antibodies is further subdivided into:

- Predicted tractable (mid to low confidence)

- Predicted tractable (high confidence)

- Clinical precedence

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 to 343. Let's focus on ERBB2 by searching for this gene symbol:
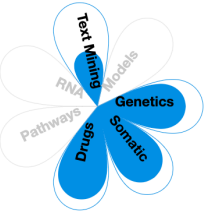
Cancer Gene Census selected

Search for ERBB2

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

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

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

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

Genetic associations
Somatic mutation
Drugs
Text mining

Grey areas in the flower plot above indicate there is no information for the corresponding data types.

Let's now scroll down on the page and expand the tabs available, for example, Genetic associations (with the browser view), Somatic mutations (with links to COSMIC) and Drugs (with links to clinicaltrials.gov). There are currently four drugs currently listed in different phases of different clinical trials targeting ERBB2 in patients with ovarian carcinoma (or children terms of this disease). Do you know why there are nine entries, although only four unique drugs? Which trials are in the advanced phase III?

Have a look at the results from our text mining approach as well and see which papers have been used as evidence for the associations between ERRB2 and ovarian carcinoma.

Once we explore the evidence and go through the different data listed, we can move on to look at the annotations for the target (ERBB2) and disease (ovarian carcinoma) in the target profile and disease profile pages, respectively.

Demo 2: Target and disease annotations

Let's now scroll back up to the top of this 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 gene:

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

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, and cancer hallmarks:

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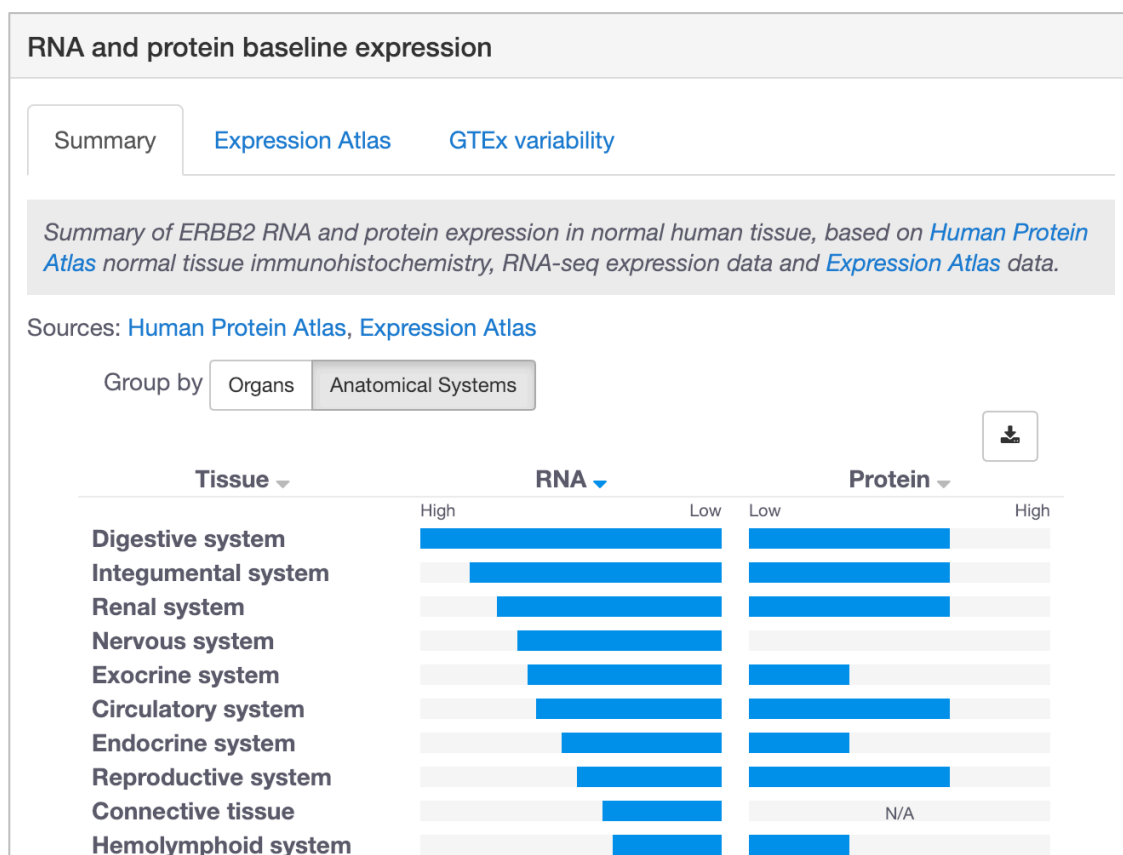
Note that from the target profile page you can also explore other diseases associated with ERBB2, apart from ovarian carcinoma, if you click on "View associated diseases":

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

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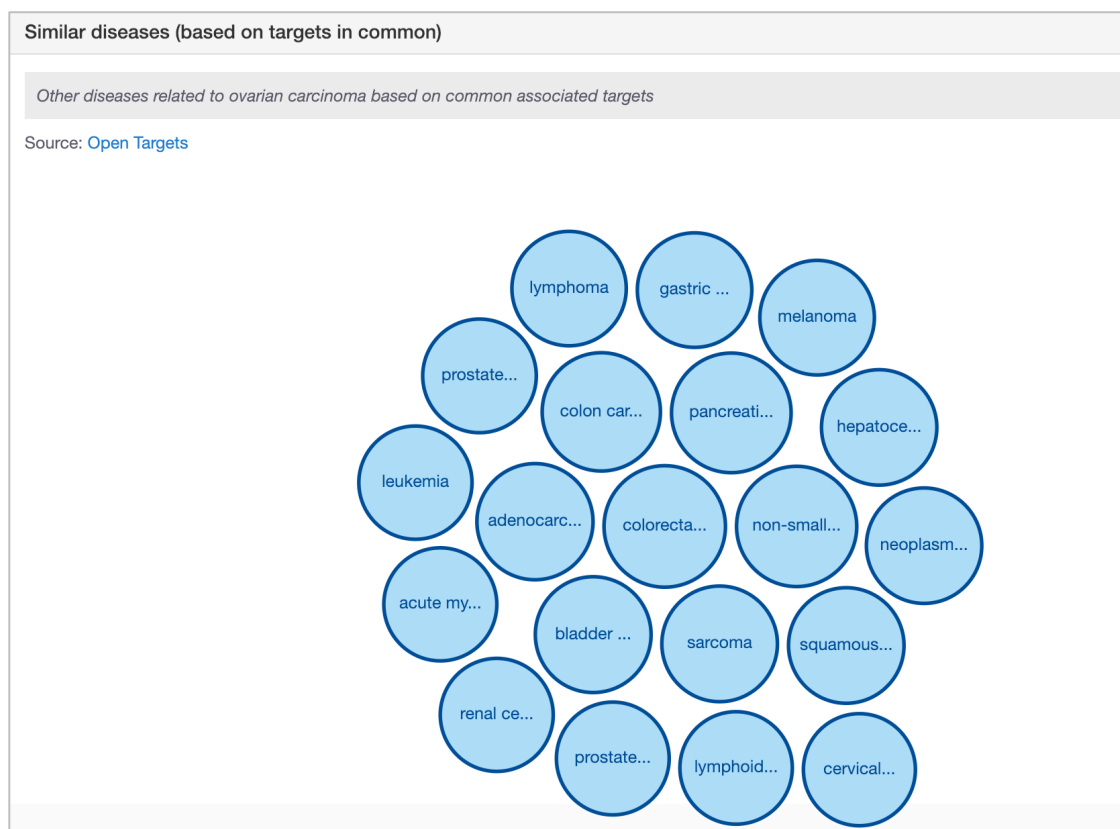
For each pair of diseases, we compute the overlap of shared targets against the total number of connections to both targets, correcting each pair by the significance and the specificity of these connections.

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.

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):
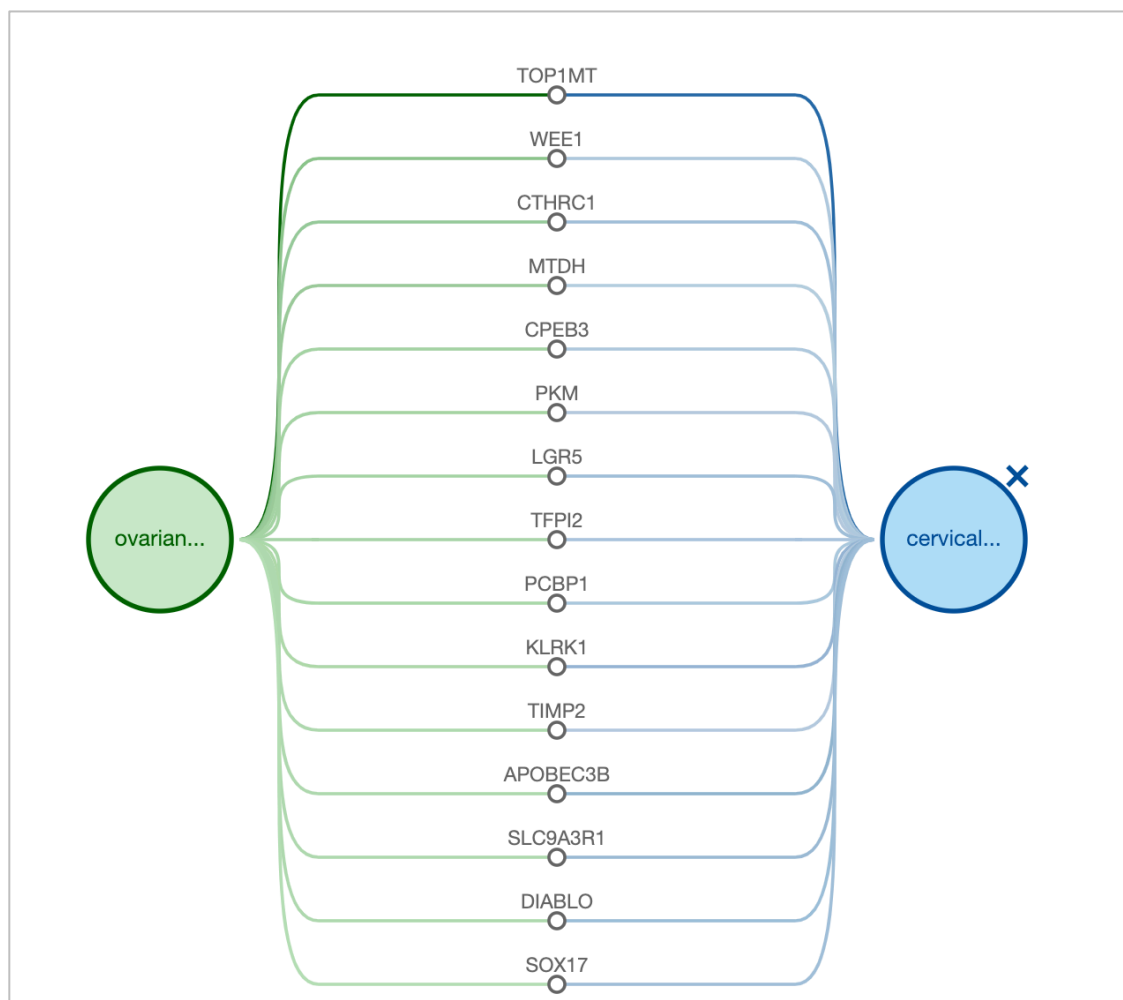
Demo 3: Target centric workflow

You can also use the Open Targets Platform from a target standpoint. 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:
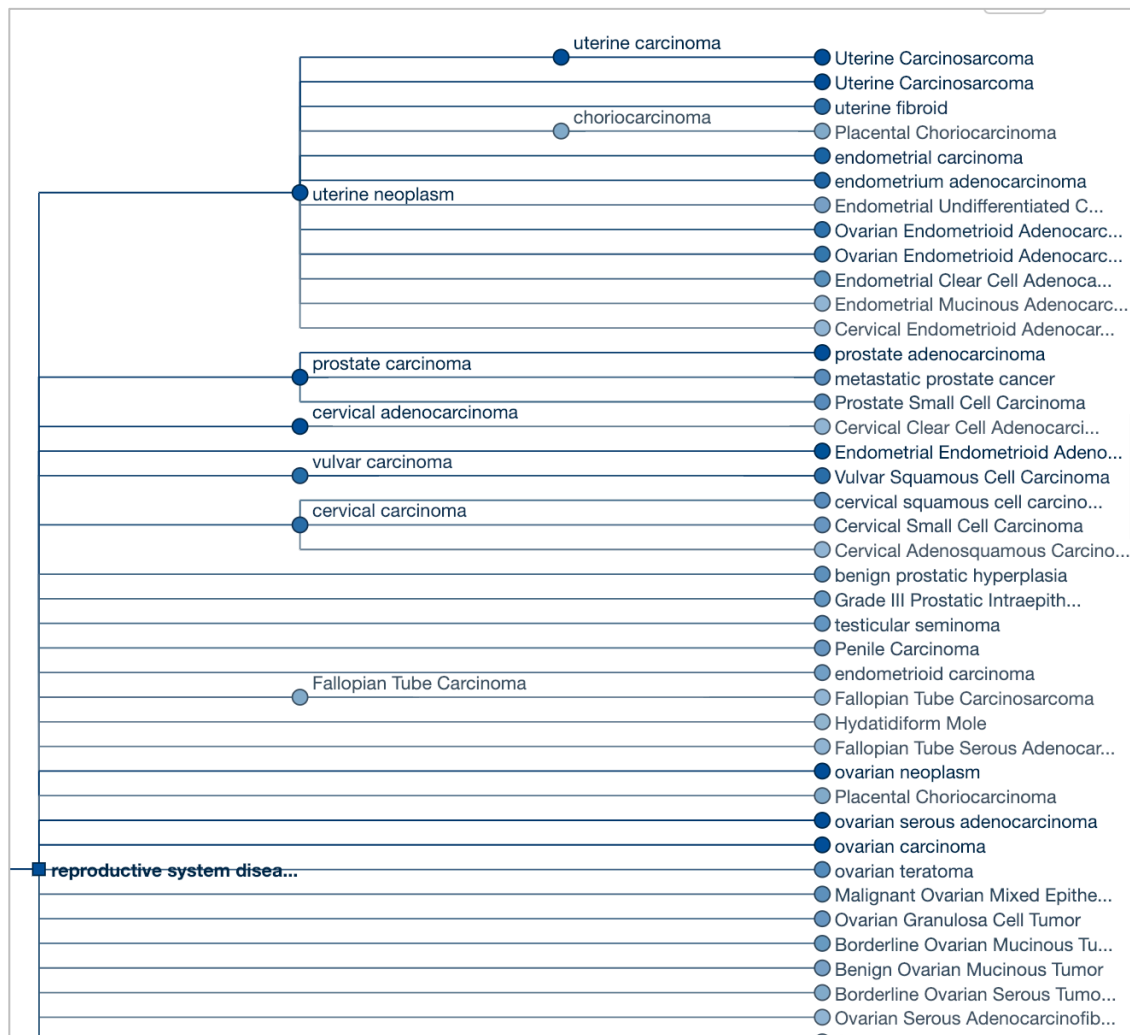
For all these three different views, you have the option to filter the data according to Data type or Therapeutic area. You can, for example, explore all diseases of the Reproductive system associated with AR. In the April 2019 release of the Open Targets Platform, there are 112 diseases of the Reproductive system disease, such as 'cervical adenocarcinoma' and 'prostate adenocarcinoma'.

You can view the same information in a bubbles 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:
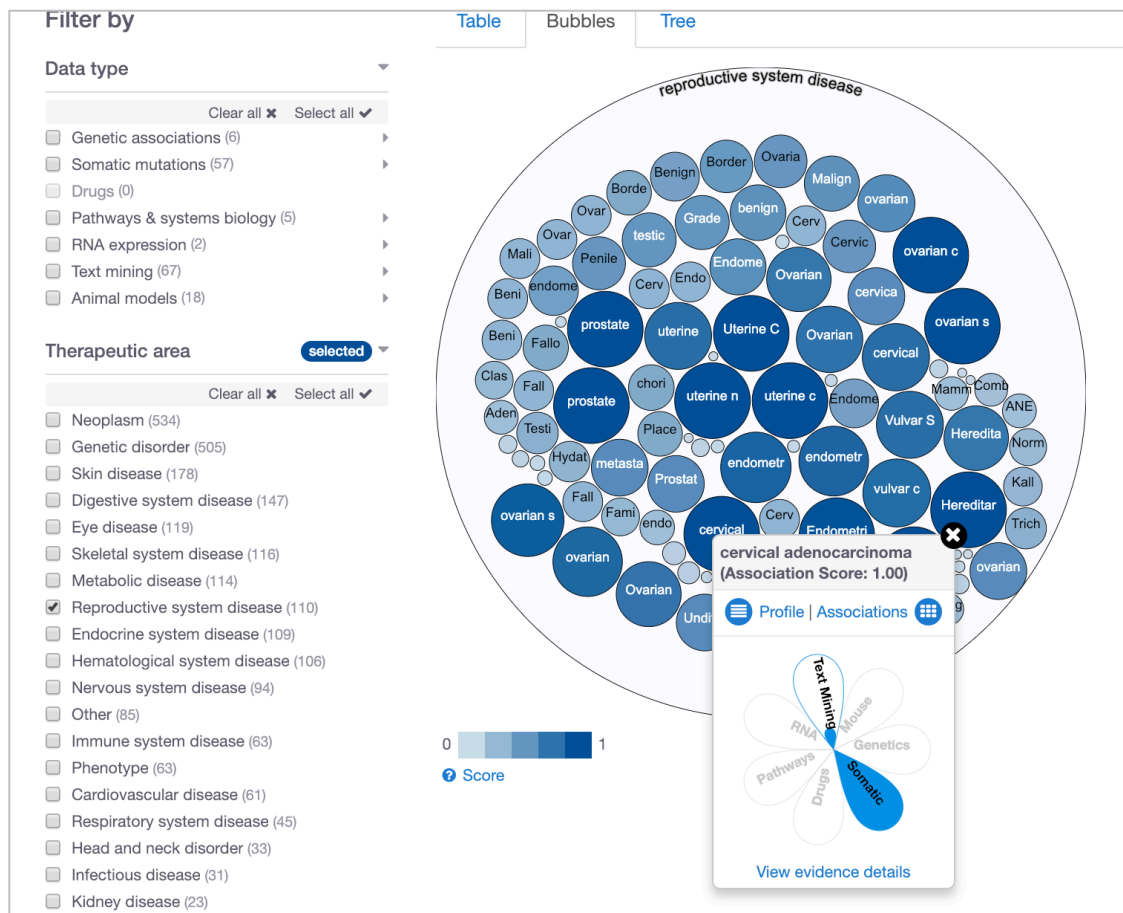
- "Profile" to get to the profile page of the disease

- "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.

Note that in the associations page for AR (or any other target), you can click on "View AR profile" to see the annotations for this gene that can help you to prioritise which targets to pursue. Check the new CRISPR-Cas9 annotation and Target safety, for example.

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 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? Note the extent of the match between search term (durvalumab) and the hits returned on the search results page.

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

c) Is there any evidence showing a decrease in RNA expression of CD24 in studies comparing primary tumour tissue from non-small cell lung carcinoma and adjacent normal tissue? A decrease in RNA expression in the tumour tissue would suggest the gene is down-regulated in the disease state.

d) Let's now explore some annotations for this disease (this is available in the disease profile page). Are there other drugs than durvalumab in phase III of clinical trials where the status is "recruiting" patients?

e) Now click on the durvalumab link in the Drug information table of the disease profile page of non-small cell lung carcinoma. What is the mechanism of action of this drug? Have any adverse effects been reported for this drug? When was this drug first approved? In

addition to non-small cell lung carcinoma, can you name other diseases where this drug is in clinical trials for?

Exercise 2: Advancing research in the field of IBD

BACKGROUND
More than five million people worldwide live with IBD, whose symptoms can be unpredictable. While the causes of IBD are unknown, several hypotheses have been suggested so far including genetic predisposition, environmental triggers, chronic and aberrant inflammation. Go to the Open Targets Platform to answer the following:

QUESTIONS
a) How many targets associated with IBD are there, which proteins are involved in the interleukin-4 and 13 signaling pathway (immune system) and for which there is GWAS catalog evidence?

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

b) Let's focus on this target and explore its profile page to answer the following:

- Look at the target tractability information for this target. Is there a higher confidence to identify a small molecular or an antibody to modulate this target?

- Which amino acids correspond to antigenic sequences in this protein where an antibody could bind to? Hint: Expand the Protein information tab).

- In addition to Interleukin-4 and Interleukin-13 signaling, what other pathways is this protein involved in?

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

Exercise 3: WRN, a promising target in MSI tumours

BACKGROUND
Our recent paper [Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens](#)) has revealed WRN as a potential drug target that could be modulated by antagonists as adjunct therapy to approved immune checkpoint inhibitors in MSI tumours.

QUESTIONS
a) How many diseases in the therapeutic area Neoplasm are associated with WRN?

b) One of these associations is ovarian carcinoma. Which evidence has been used for its association with WRN?

c) Can you determine if the somatic mutation (s) used for the WRN-ovarian carcinoma association is/are direct or indirect evidence?

d) Finally, have a look at the text mining evidence from both full text and/or abstract in research articles. Can you find the sentences in the paper (s) where target and disease names co-locate?

Exercise 4: Filtering Alzheimer's disease associations based on a list of targets

BACKGROUND
A drug discovery scientist at Alzheimer's Research UK has a list of eight targets that seem to be associated with Alzheimer's disease (AD) based on literature reviews.

Can you find out the information Open Targets has for these genes?
HFE
PSEN1
PRO1557
APOE
ADRB2
PSEN2
CPAMD5
BACE1

QUESTIONS

a) How many of these targets have data on animal models used for the association with Alzheimer's? Is the association supported by direct or indirect evidence?

b) Let's go back to the associations page listing all eight targets from your list. Which of these eight targets have higher levels of mRNA expression in the cerebral cortex than in any other tissue (this is known in Open Targets as RNA tissue specificity)?

c) Let's now explore the disease profile page of Alzheimer's. Can you name a few diseases that are similar to Alzheimer's based on targets in common?

EXTRA HANDS-ON EXERCISES

Exercise E1: Assessing the specificity of a list of targets for Barrett's esophagus

BACKGROUND
A biologist working on translation medicine at the Imperial College of London has a list of genes linked* to Barrett's esophagus but he/she would like to know how specific this list to that disease or whether this set of genes could be therapeutic targets for other diseases of the digestive system.

QUESTIONS
a) Are there other diseases affecting the esophagus where this list of targets is also specific to?

b) What is the most enriched pathway for this list of genes?

c) Which of those genes seem to be the most tractable ones for either small molecule or antibody.

d) Are there any drugs currently in clinical trials phase I or II targeting any of these genes?

e) Is any protein-protein interaction predicted to exist in this set of genes?

*The list can be download from https://tinyurl.com/batch-kogo-0219.

Exercise E2: Retrieve associations with the REST-API

BACKGROUND
So far you have used the user interface of the Open Targets Platform and the batch search. Other modes of data access include the Open Targets REST API and data downloads.

BACKGROUND

The following three genes have been associated with gastric carcinoma: ERBB2, TP53 and PRKAA1

QUESTIONS
a) Can you use the REST API to find all diseases (besides gastric carcinoma) associated with these three genes?

b) Which diseases have the highest overall association score for each of those three genes?

c) Can you download the above list in TAB format?

d) Can you retrieve all targets associated with gastric carcinoma filtered by somatic mutations and with an association score greater than 0.9?

e) Let's now retrieve the evidence from the Cancer Gene Census for the association of gastric carcinoma and KRAS. Can you compare to the evidence you see on the user interface?

Check our help docs for more [API tutorials](#).

Exercise E3: 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) Can you list a few phenotypes related to cancer (neoplasm) that are observed when Egfr gene is knocked-out in mice?

b) Is EGFR involved in the promotion or suppression of tumour associated angiogenesis? (tip: look for the cancer hallmarks).

c) In addition to mouse, could you use other model organisms (e.g. guinea pig, rat) to assess what the modulation of EGFR would entail (hint: check the "Gene tree" tab)?

d) Under Bibliography, which papers published in 2019 are available for EFGR and the following combination of additional terms: "cetuximab", "treatment" and "head and neck squamous cell carcinoma"? Pick any of these papers and click on "Show abstract" to explore the annotations available. You may also want to explore the similar articles available for any given paper.

Exercise E4: LRRK2 in Parkinson's disease

BACKGROUND
The LRRK2 gene encodes a protein with five putative functional domains: an N-terminal leucine-rich repeat (LRR) domain, a Roc (Ras of complex protein) domain that shares sequence homology to the Ras-related GTPase superfamily, a COR (C-terminal of Roc) domain, a mitogen-activated protein kinase kinase kinase (MAPKKK) domain, and a C-terminal WD40 repeat domain. A genetic variant in this gene is one of the most common causes of inherited Parkinson disease (Gandhi et al., 2008).

QUESTIONS
a) How long is the protein encoded by this gene/target? Can you find the protein domains listed above?

b) No drug is currently available to target LRRK2. There may be other compounds, such as chemical probes (small molecules that will alter the function of biological target)? Can you use the Open Targets Platform to see which chemical probes, if any, are available that could be used to modulate the function of this protein?

c) Can you look at the probability of this protein to be targeted by either a small molecule or antibody (tip: have a look at the Target tractability information).

d) Can you list some of the proteins that interact with LRRK2? Can you download this image?

e) Let's now have a look at the diseases associated with this target. Can you download a table containing all diseases from the nervous system where there is evidence for the associations from the GWAS Catalog?

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 (Schubert et al) 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) 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)

- 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.

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