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

Hands-on Workshop
Answer booklet

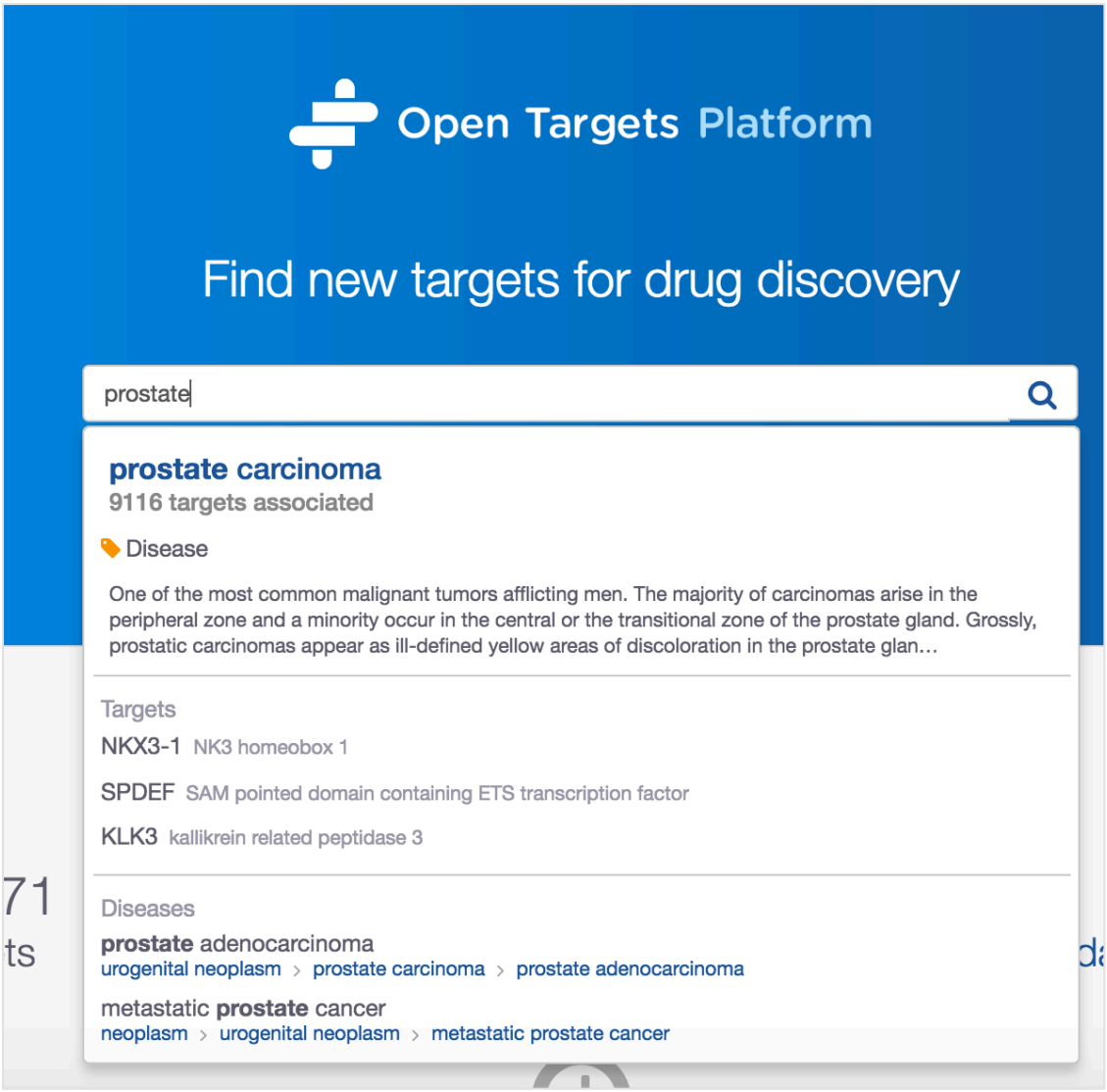
CRUK Cambridge Institute
16th December 2016

Denise Carvalho-Silva
Open Targets Outreach

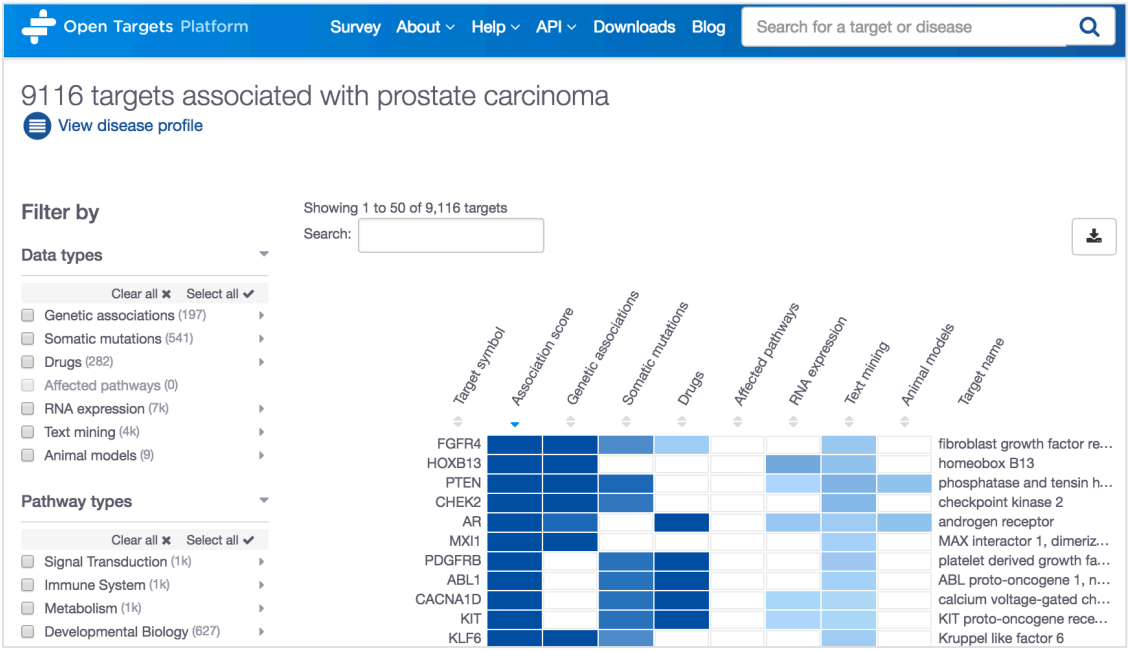
Answers to exercises 1 and 2, pages 25-27 of coursebook

Exercise 1 – Prioritising targets for drug discovery in prostate carcinoma

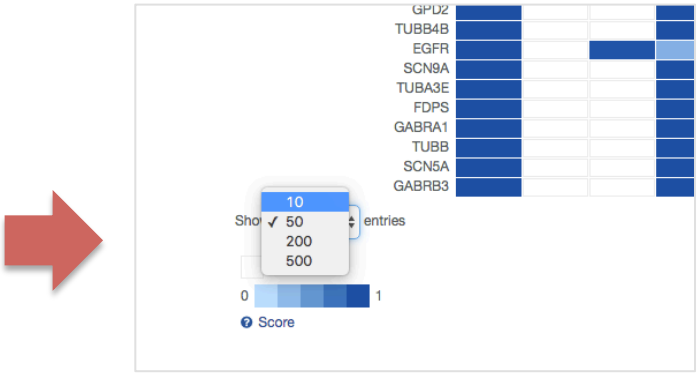
a) Go to www.targetvalidation.org and search for *prostate carcinoma*:



Select the first (best) hit. You will then see a page like this, which lists 9116 targets associated with prostate carcinoma:



Scroll down and select to see the results with 10 entries (rows) only:



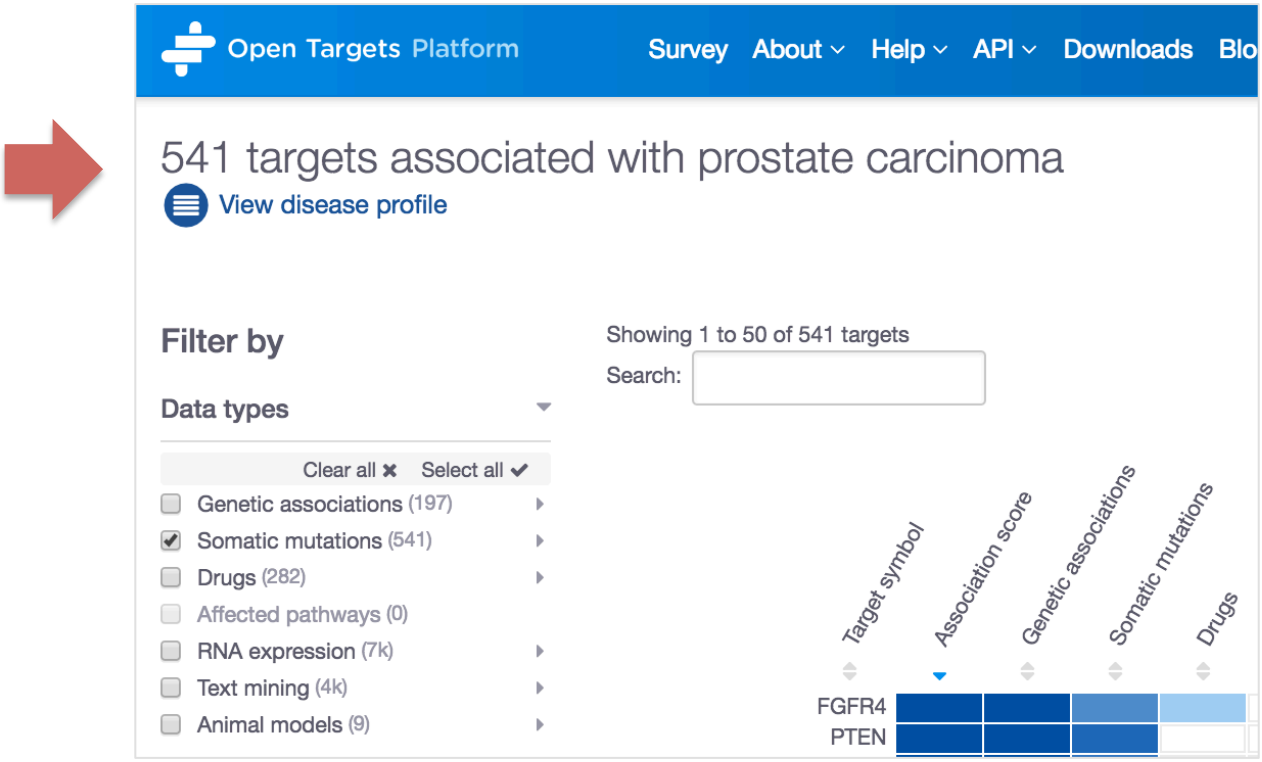
The first 10 rows will show the top 10 targets associated with prostate cancer. These will have the highest score (score of 1): *FGFR4, HOXB13, PTEN*, *CHEK2*, *AR*, *MXI1*, *PDGFRB, ABL1*, *CACNA1D*, and *KIT.*

The confidence on the target-disease association is indicated by the association score, which ranges from 0 to 1 (from no association to the strongest association).

The score is computed individually for each piece of evidence (e.g. a drug on phase I), followed by the score computed for the data sources (e.g. ChEMBL), then a score for the data type (e.g. Genetic associations) and the overall score (a harmonic sum of the individual scores). The overall score is shown in the first column of the table shown above. More details on the scoring can be found below:

https://github.com/CTTV/association_score_methods
nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

b) Restrict the results by filtering the table to show the targets associated with prostate cancer based on Somatic mutations only:



This filtered list (restricted to somatic mutations only) shows different targets than the list resulting from step (a) above.

There are no somatic mutations described in the *HOXB13* and *MXI1* genes, therefore *HOXB13* and *MXI1* are no longer in the top 10 targets based on somatic mutations.

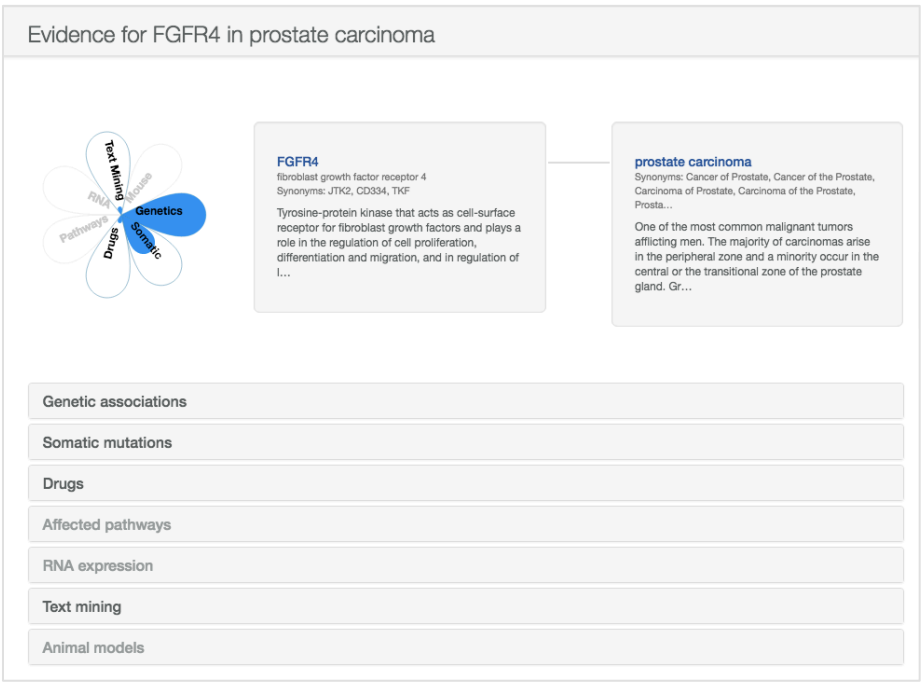
For more details on the data we currently use to associate a gene to a disease can be found below:

https://www.targetvalidation.org/data_sources

c) Let's now focus on one of these targets namely *FGFR4* to find out more about some of the evidence that seems to support the association between *FGFR4* and prostate carcinoma.

Click on the gene name itself or on any cell in the gene table that corresponds to the *FGFR4* row:

This will take you to a page similar to this:

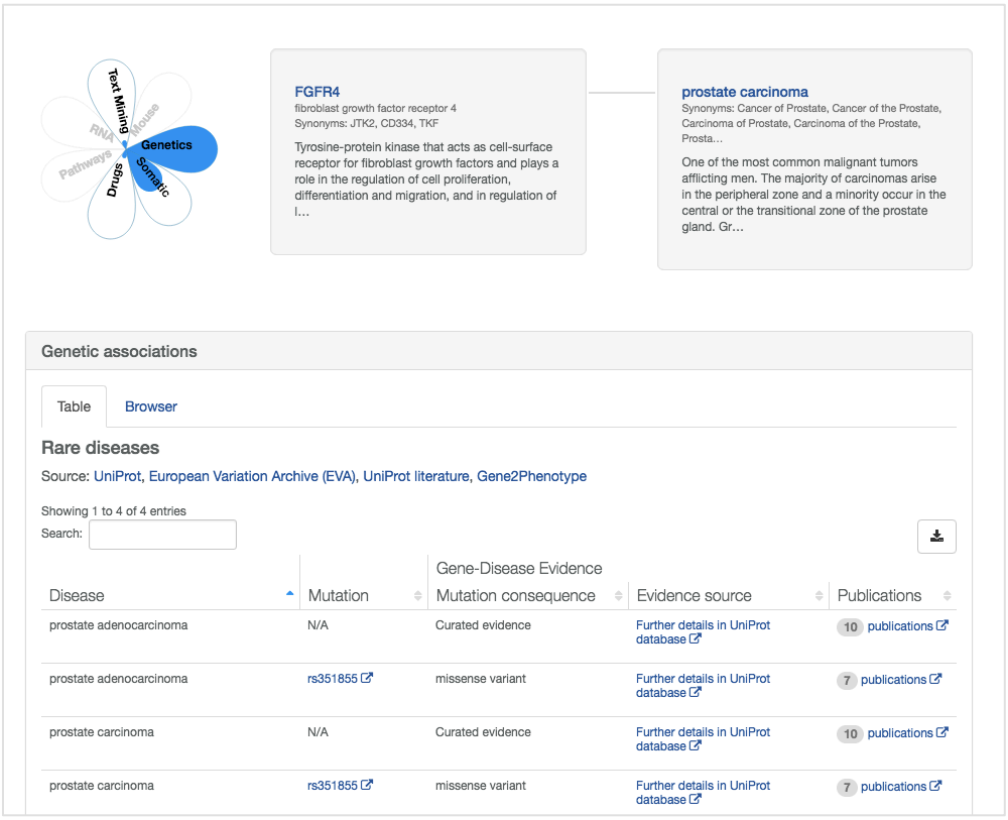


The evidence used to support the association is shown in different tabs (the grey tabs have no data: there is no data for Affected pathways, RNA expression and Animal models to support FGFR4-prostate carcinoma association).

Expand the 'Genetic associations' tab.

Tip: if you click on the cell containing the data relative to Genetic associations (see below):

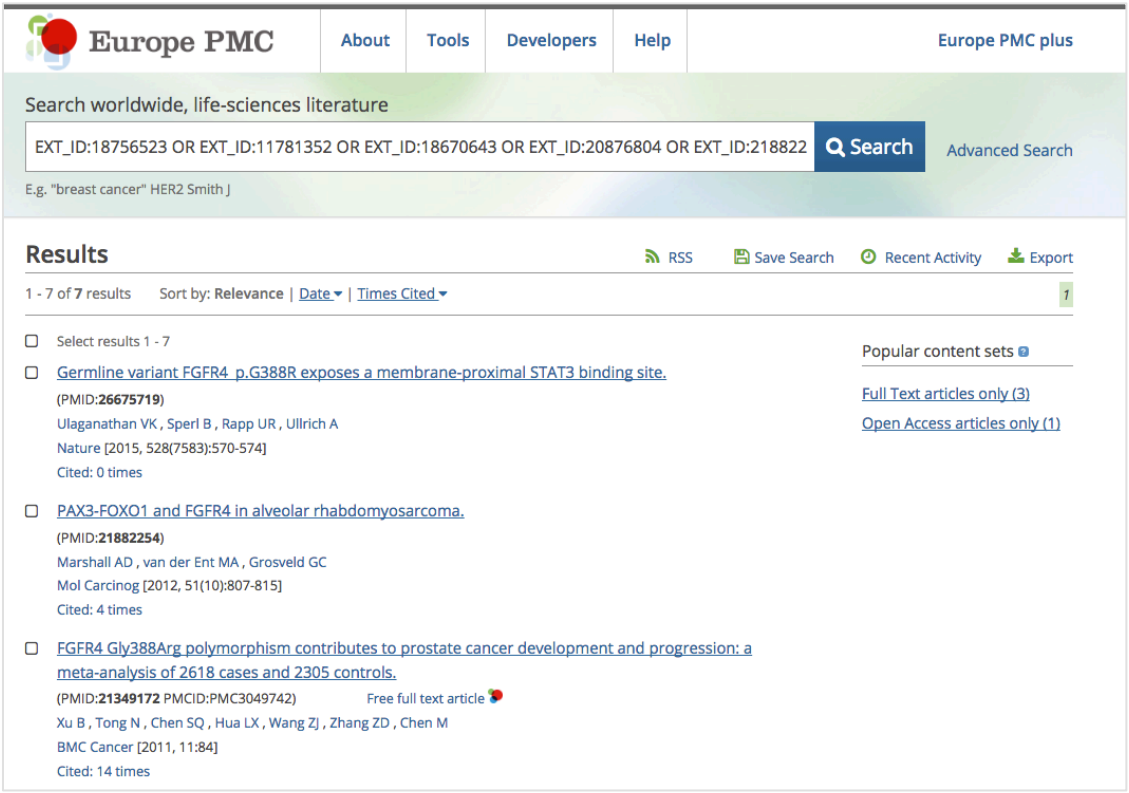
you will automatically land on a page where the tab containing the Genetic association will be already opened:



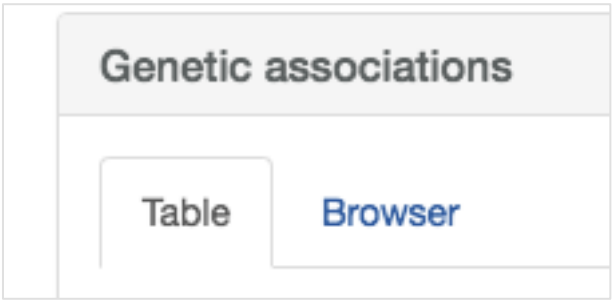
Yes, there is one genetic variant that is known in public databases to be associated with prostate carcinoma. Its ID is rs351855.

Note that we aggregate evidence from highly specific terms of the disease ontology (e.g. prostate adenocarcinoma) to broader, parent terms (e.g. prostate carcinoma).

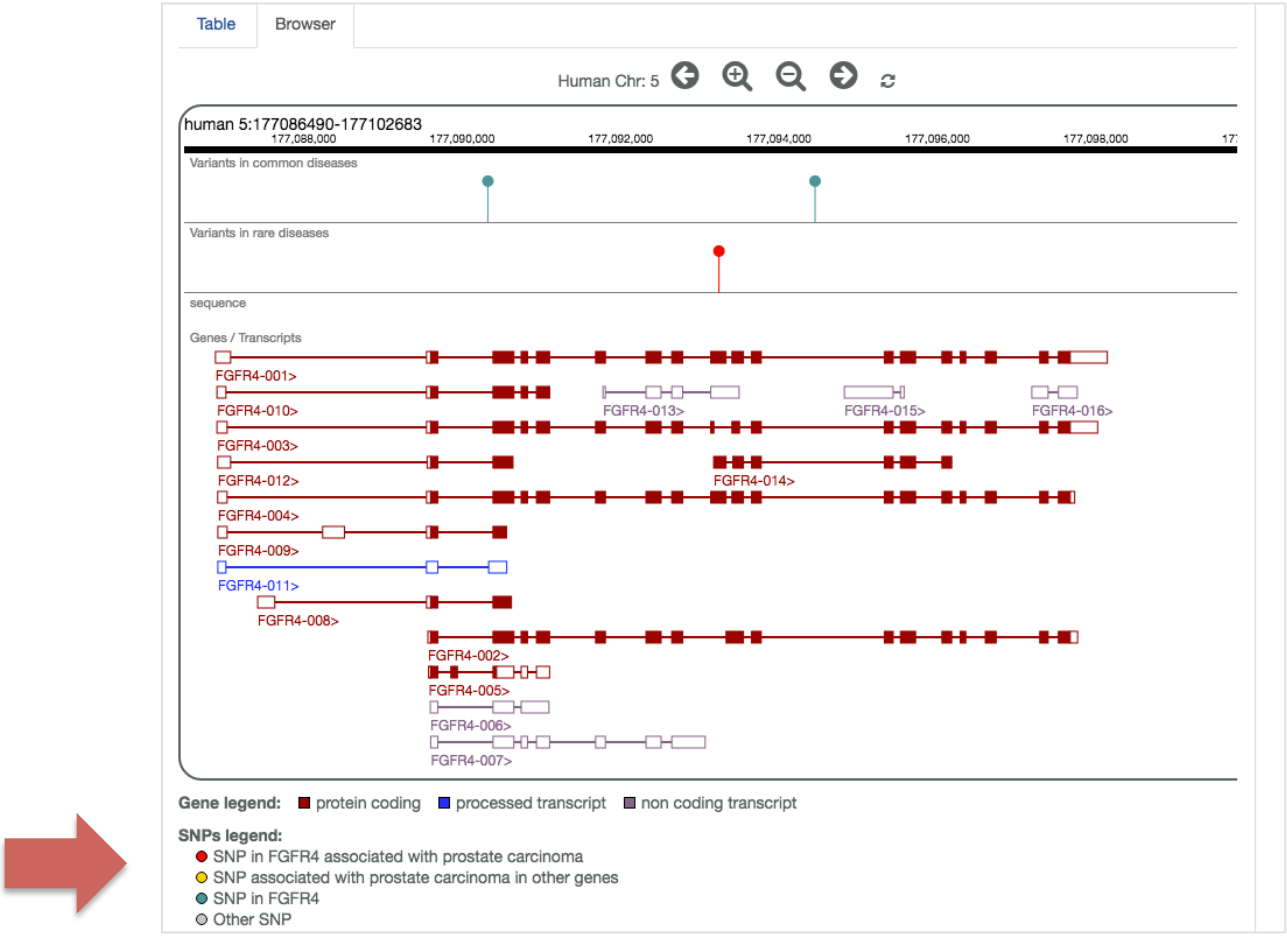
Click on the 7 'publications' link to see the papers supporting the association:



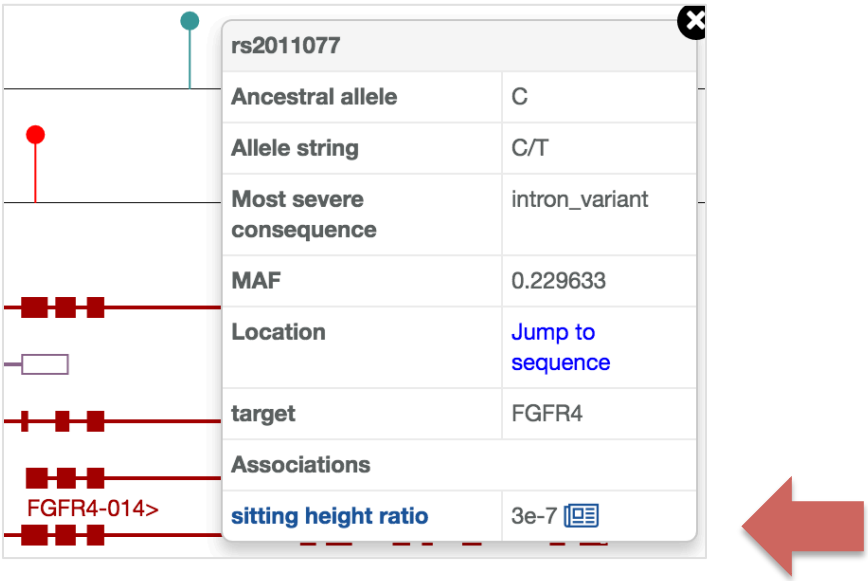
d) The Genetic associations data can also been visualised in a graphical display. Click on the 'Browser' link:



You will see the transcripts mapped to that gene and the variants (SNPs or mutations) in the region. Check the legend to find out what the colours mean. This browser view is interactive and dynamic: you can zoom in and out and scroll along the genomic region.

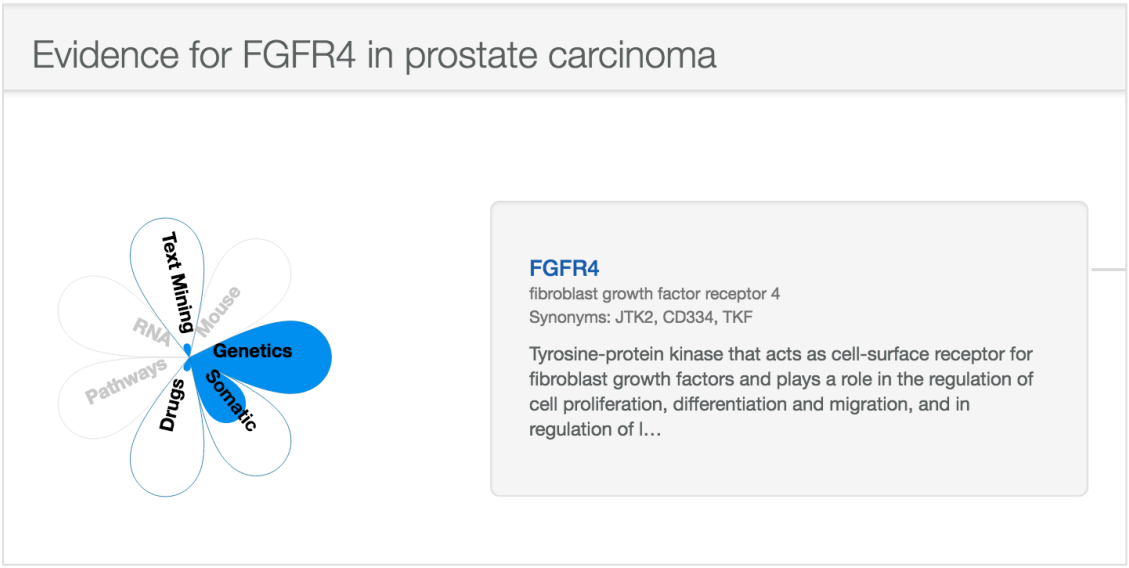


In this genomic region, you can see there are two mutations associated with other traits (i.e. sitting height ratio and body height). You may want to zoom out to view more variants up or downstream of the gene, and then click on the lollipop (the variant) for more details:



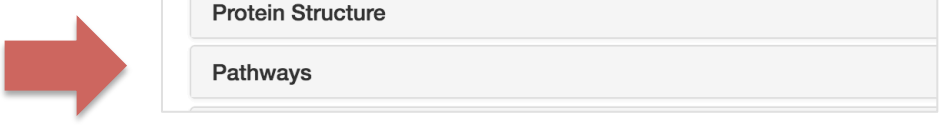
e) Let's now have a look at the target itself. This will give us information outside the context of any disease.

Still on the same page as above, click on the hyperlink FGFR4:



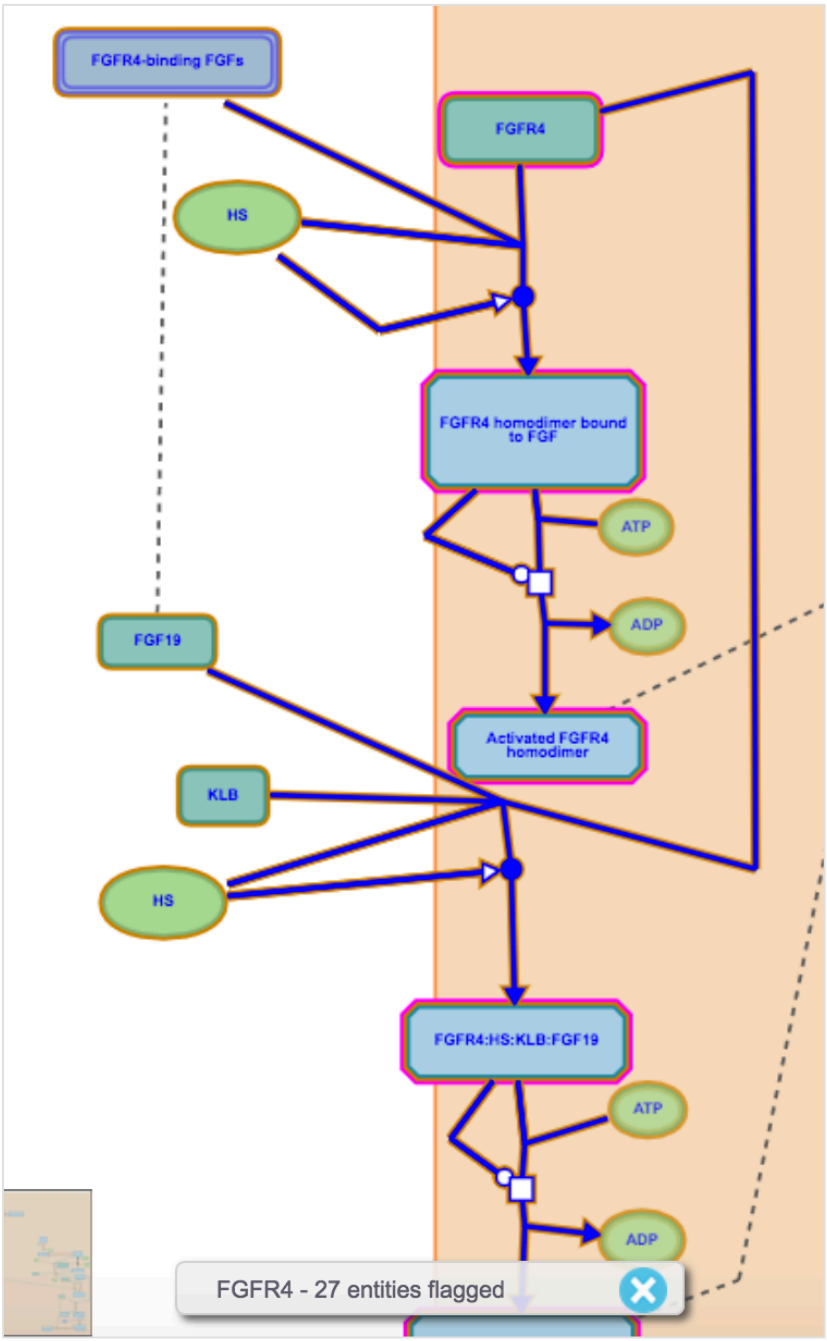
You will end up in a page like this:

https://www.targetvalidation.org/target/ENSG00000160867

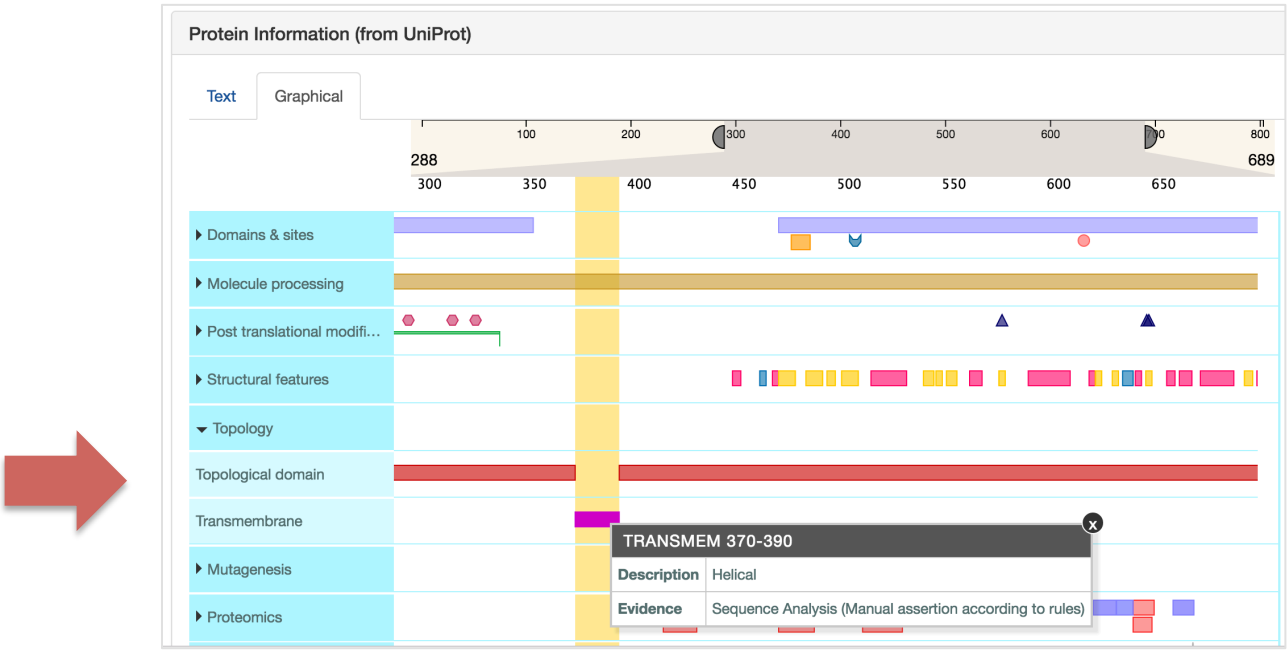


Click on Pathways to find out cellular pathways and biochemical process this gene is involved in e.g. PI3K Cascade, Constitutive Signaling by Aberrant PI3K in Cancer and few others.

You can visualise FGFR4 ligand binding and activation pathway in an interactive display:



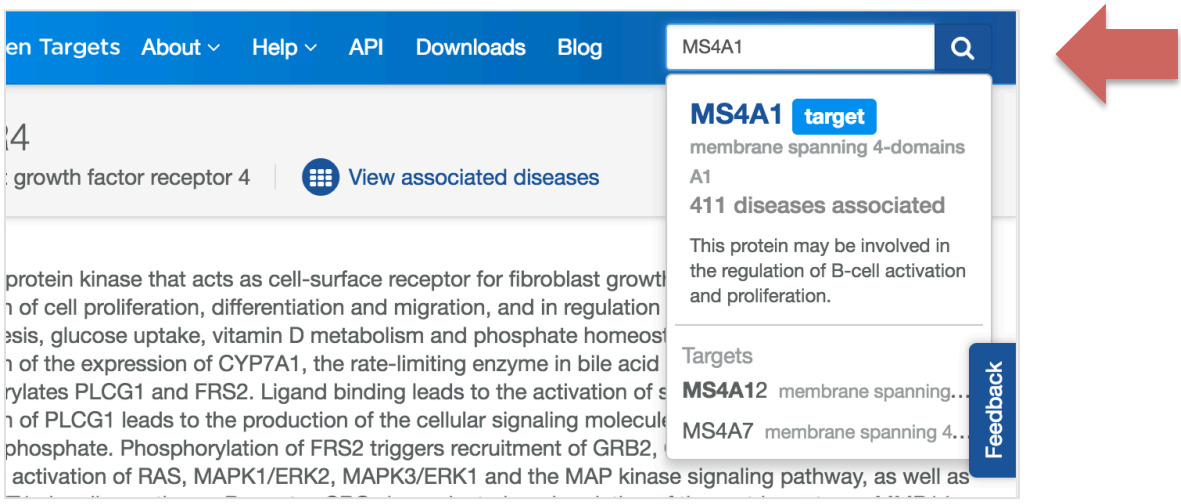
f) Click on the Protein information (from UniProt) tab. Now, click on the Graphical view option, then click on the Topology menu to see the annotated domains: extracellular, transmembrance and intracellular. The transmembrane (TM) domain goes from amino acid 370 to 390. Gene *FGFR4* codes for a receptor, so one should expect a transmembrane domain to be annotated in the protein.



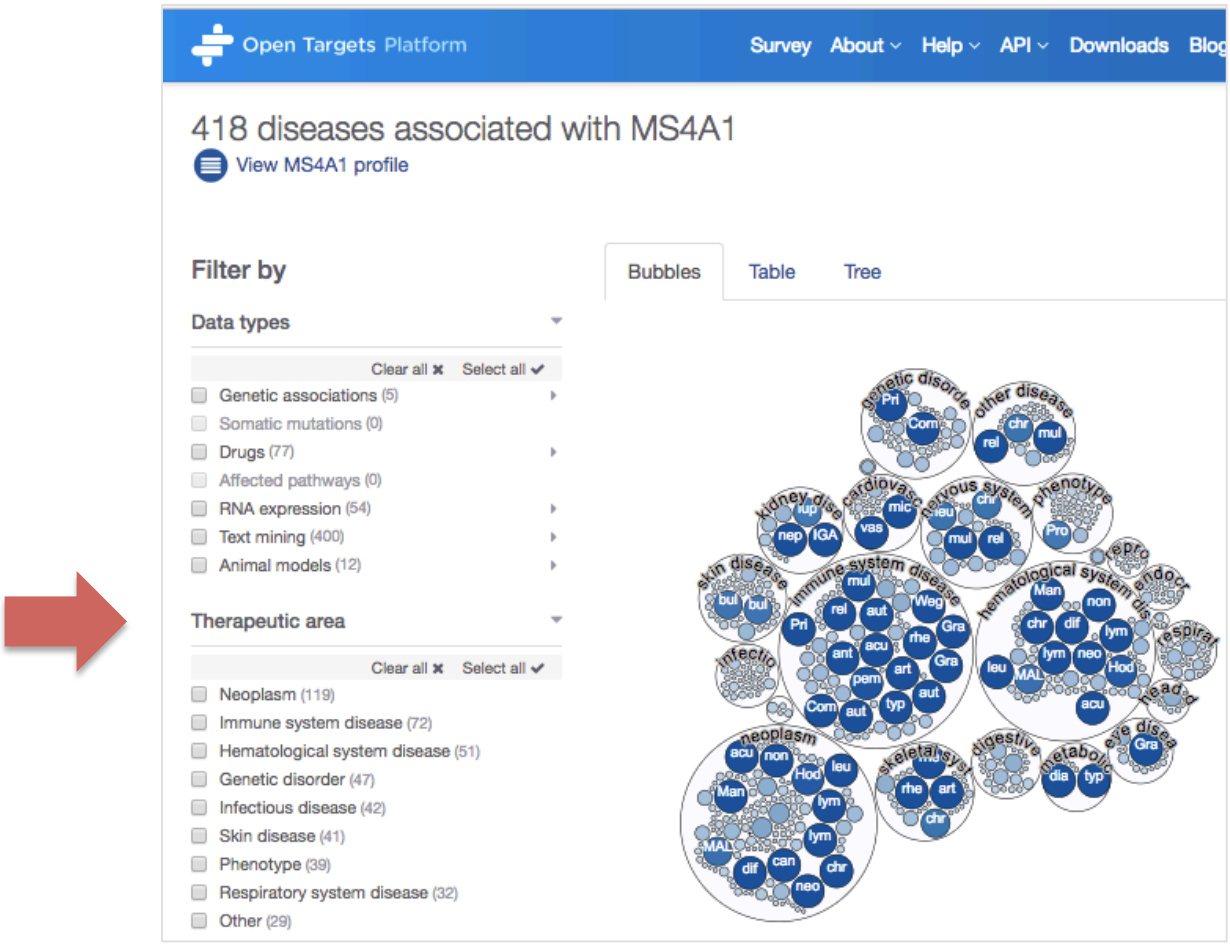
Exercise 2 – *MS4A1* as a possible drug target in the treatment of non-Hodgkin's lymphoma

Search for MS4A1.

Tip: you do not need to go back to the homepage: you can use the search box at the top right corner of any pages in the Platform:

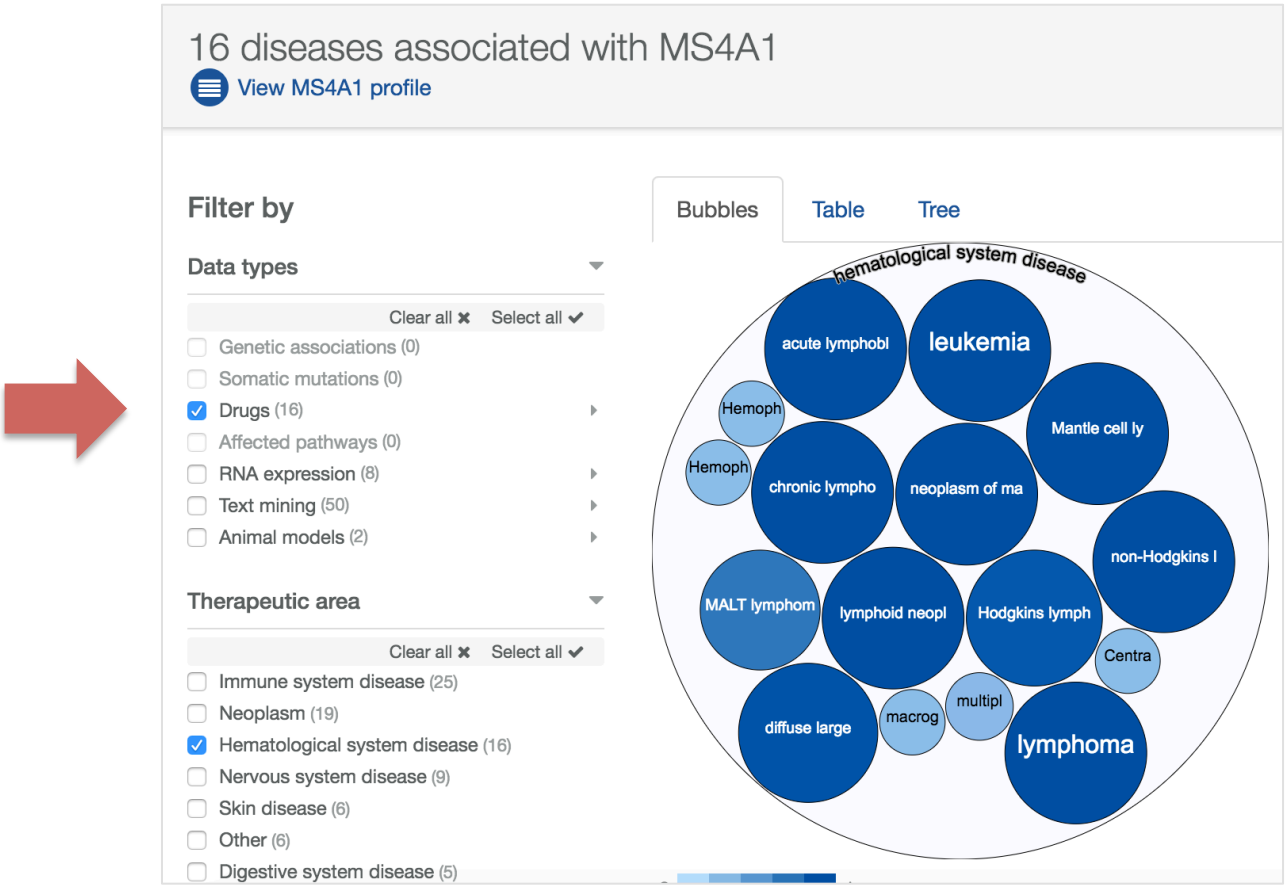


There are 418 diseases associated with target *MS4A1*.



You can filter the results by Therapeutic area, such as ‘Hematological system’ (which includes non-Hodgkin's lymphoma) and by Data type such as ‘Drugs’. The number of diseases associated in the

Hematological system associated with *MS4A1* for which there is Drug information is 16:



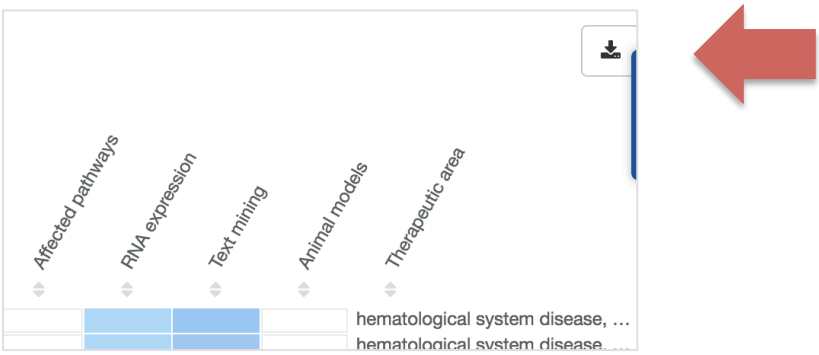
b) The default display for the diseases associated with a target is the Bubbles view.

You can view the same information as a Table or Tree.

Tip: Not sure what those views mean? Check our help page: https://targetvalidation.org/about#target_assoc

Diseases with the (overall) association score of 0.90 or above (for the filters selected) are neoplasm of mature B-cells (score of 1), chronic lymphocytic leukemia (score of 1), and Hodgkins lymphoma (0.92).

The results displayed in a Table format can be downloaded as CSV (comma separated value) and opened up in Excel:



Look for the 'Download' icon:

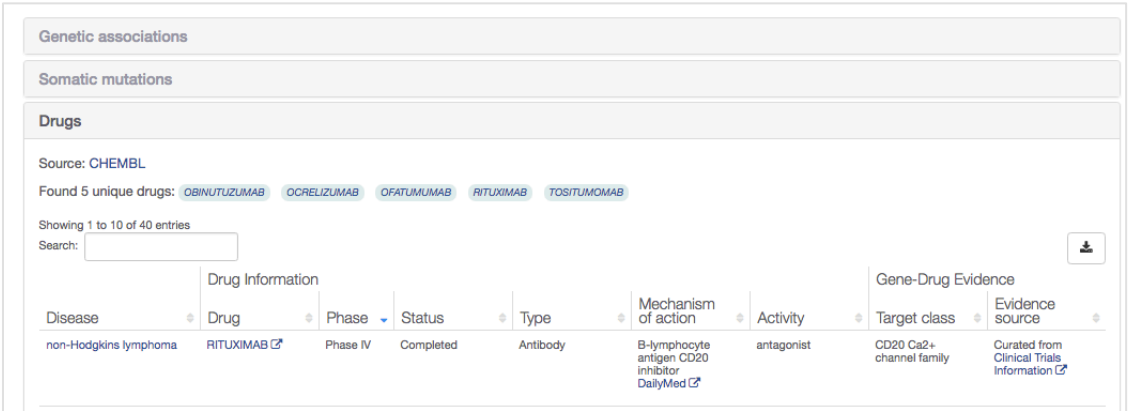
Download the table in CSV.

c) Filter the table with 'non-Hod' for 'non-Hodgkin's lymphoma'. Text mining also supports the association between the disease and *MS4A1*.

Note: the data coming form mining the literature is given a lower weight in our analysis; therefore it gets a lower score (than Drugs coming from ChEMBL for example).

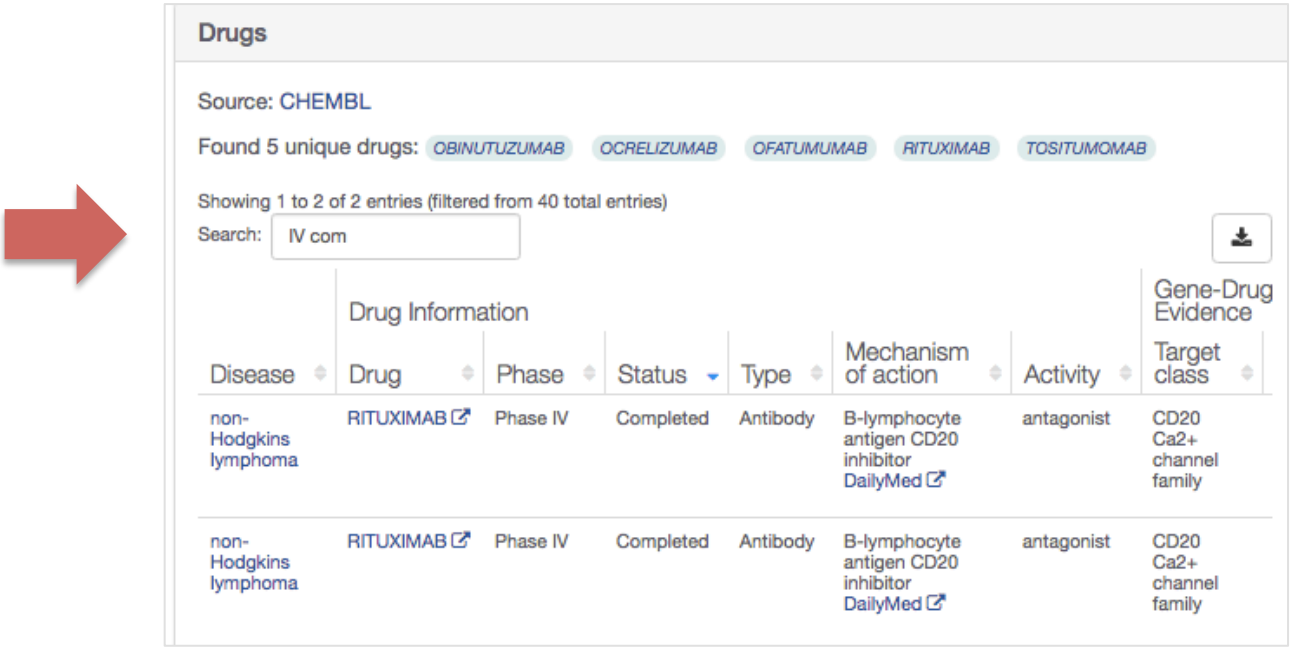
Click on the 'Text mining' cell in the table to see the 778 research articles mined from EuropePMC. These articles are flagged if they have the co-occurrence between the gene (or its synonym i.e. CD20) and the disease name in the same sentence.

d) Still in the same page as c) scroll up and expand the option 'Drugs':



Five drugs that target and modulate *MS4A1* are currently under clinical trials for the treatment of non-Hodgkin's lymphoma.

Search for 'IV comp' to limit the number of rows and find that RITUXIMAB is the only drug currently in phase IV, status completed.



Drugs

Source: CHEMBL

Found 5 unique drugs: OBINUTUZUMAB OCRELIZUMAB OFATUMUMAB RITUXIMAB TOSITUMOMAB

Showing 1 to 2 of 2 entries (filtered from 40 total entries)

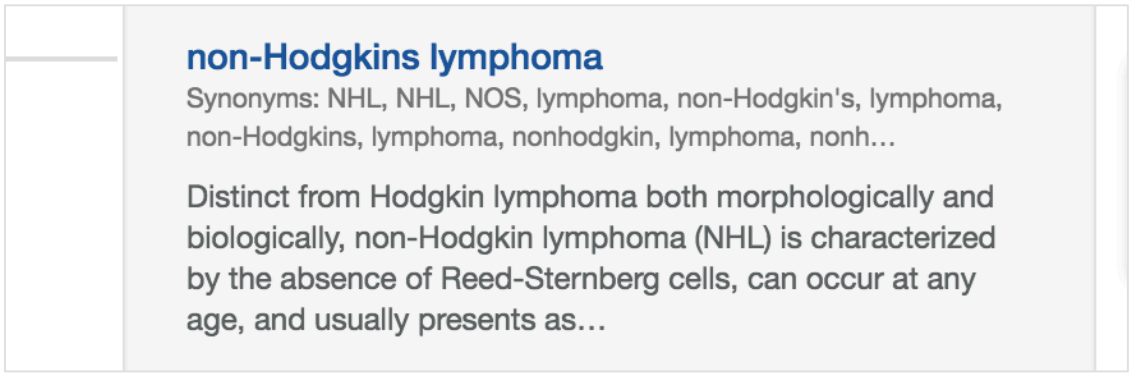
Search: IV com

	Drug Information						Gene-Drug Evidence
Disease	Drug	Phase	Status	Type	Mechanism of action	Activity	Target class
non-Hodgkins lymphoma	RITUXIMAB	Phase IV	Completed	Antibody	B-lymphocyte antigen CD20 inhibitor DailyMed	antagonist	CD20 Ca2+ channel family
non-Hodgkins lymphoma	RITUXIMAB	Phase IV	Completed	Antibody	B-lymphocyte antigen CD20 inhibitor DailyMed	antagonist	CD20 Ca2+ channel family

Why do you see two different rows with the same information in all the columns?

Some rows may look like identical at a first glance. If you click on the link in the last column (Evidence source) you will find out the drug is in phase IV completed in two different studies, NCT00090038 and NCT00430352).

e) Still on the same page, scroll up till you see the flower and click on the disease name in the box at the right hand side.



non-Hodgkins lymphoma

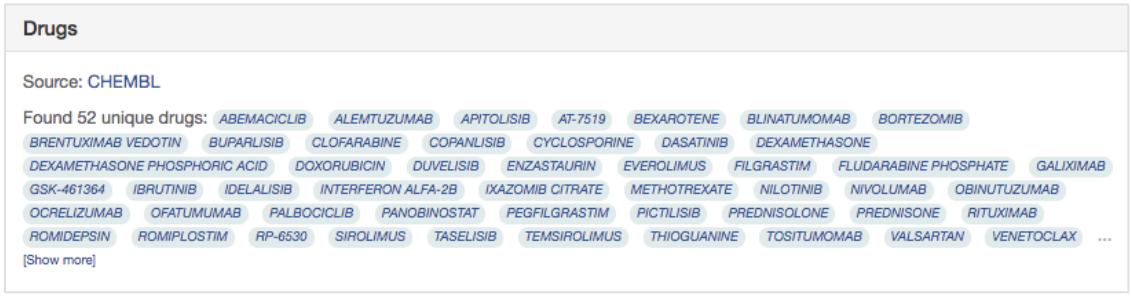
Synonyms: NHL, NHL, NOS, lymphoma, non-Hodgkin's, lymphoma, non-Hodgkins, lymphoma, nonhodgkin, lymphoma, nonh…

Distinct from Hodgkin lymphoma both morphologically and biologically, non-Hodgkin lymphoma (NHL) is characterized by the absence of Reed-Sternberg cells, can occur at any age, and usually presents as…

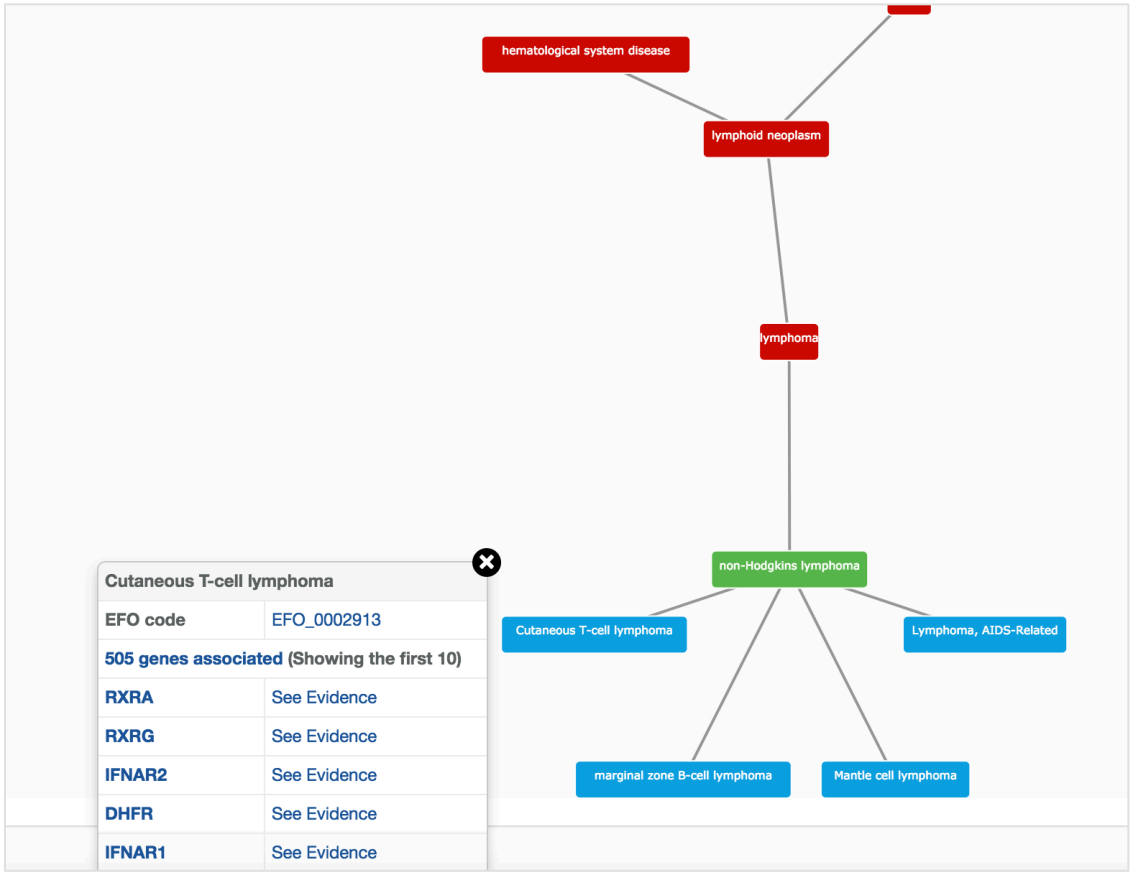
Click on the disease name above.

This will take you the disease page with more info on the disease including all drugs under investigation, its ontology and its phenotypes (if available).

The are 52 drugs linked to this disease (therefore targeting other genes, not only *MS4A1*):

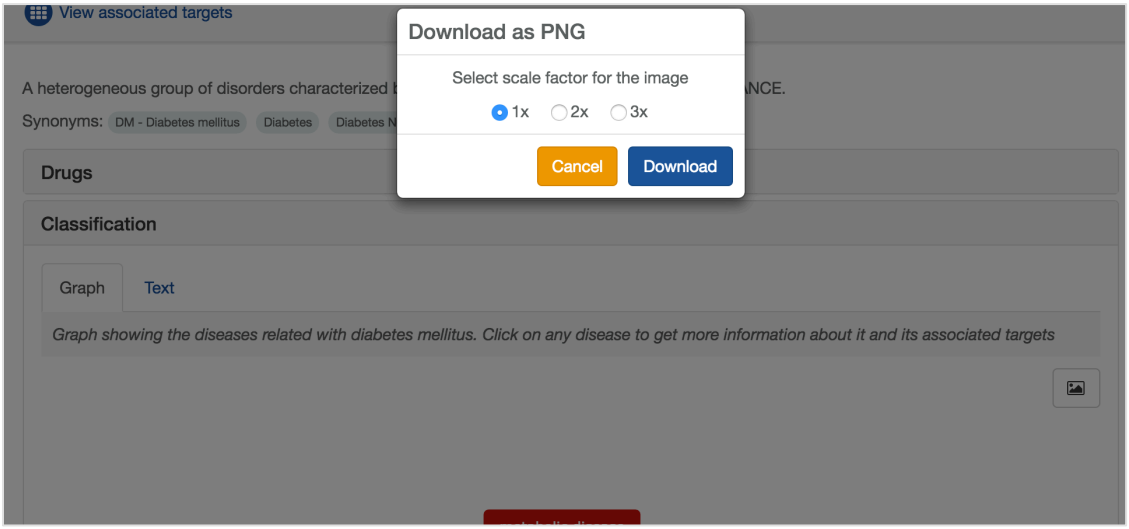


f) This is part of the diagram showing the ontology:



Click on the nodes for more information, such as the EFO ID and genes associated with the diseases in the ontology.

The disease ontology can be downloaded as a PNG format. Click on the 'Download' icon in the top right of the image:



Note some drugs you've seen in step e) have been investigated in sub-types of non-Hodgkin's lymphoma, such as Cutaneous T-cell lymphoma and Mantle cell lymphoma. This is easier to see when viewing the relationship of the parent disease (non-Hodgkin's lymphoma) and its children diseases (e.g. Cutaneous T-cell lymphoma and Mantle cell lymphoma).