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

Hands-on Workshop Coursebook

Yale School of Medicine
9th February 2017

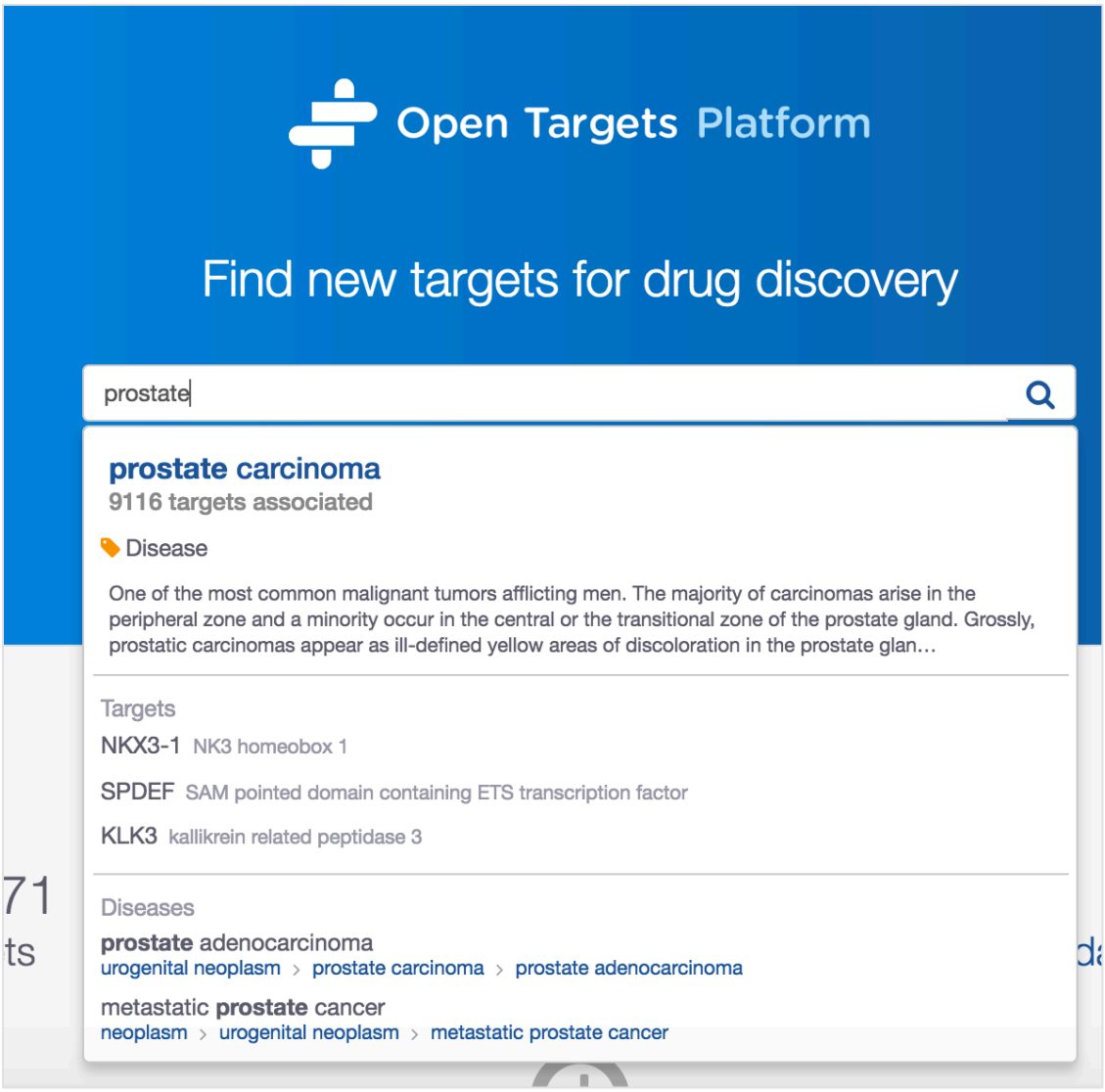
Denise Carvalho-Silva
Open Targets Outreach

Answers to exercises 1 and 2, pages 24-27 of coursebook

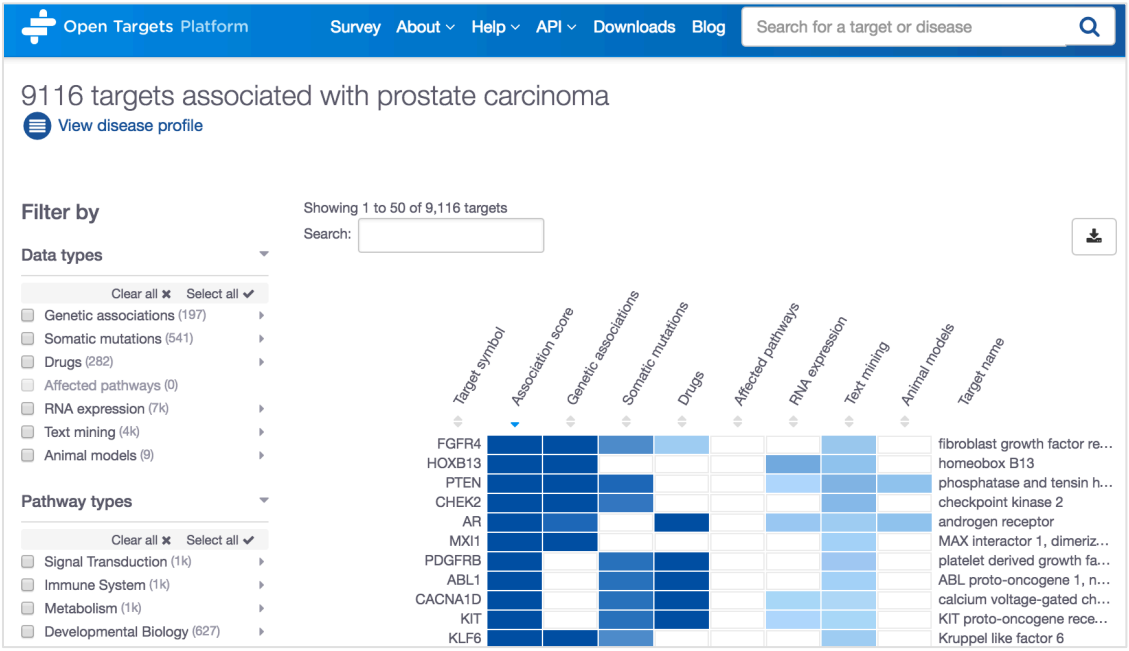
Exercise 1

Prioritising targets for drug discovery in prostate carcinoma

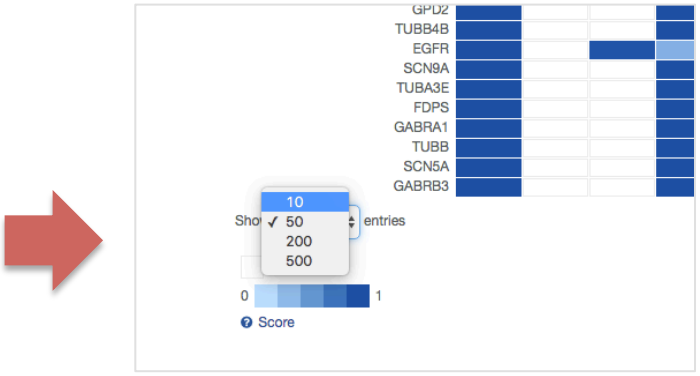
a) Go to www.targetvalidation.org and search for *prostate carcinoma*:



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 10 entries (rows) only in the result table:



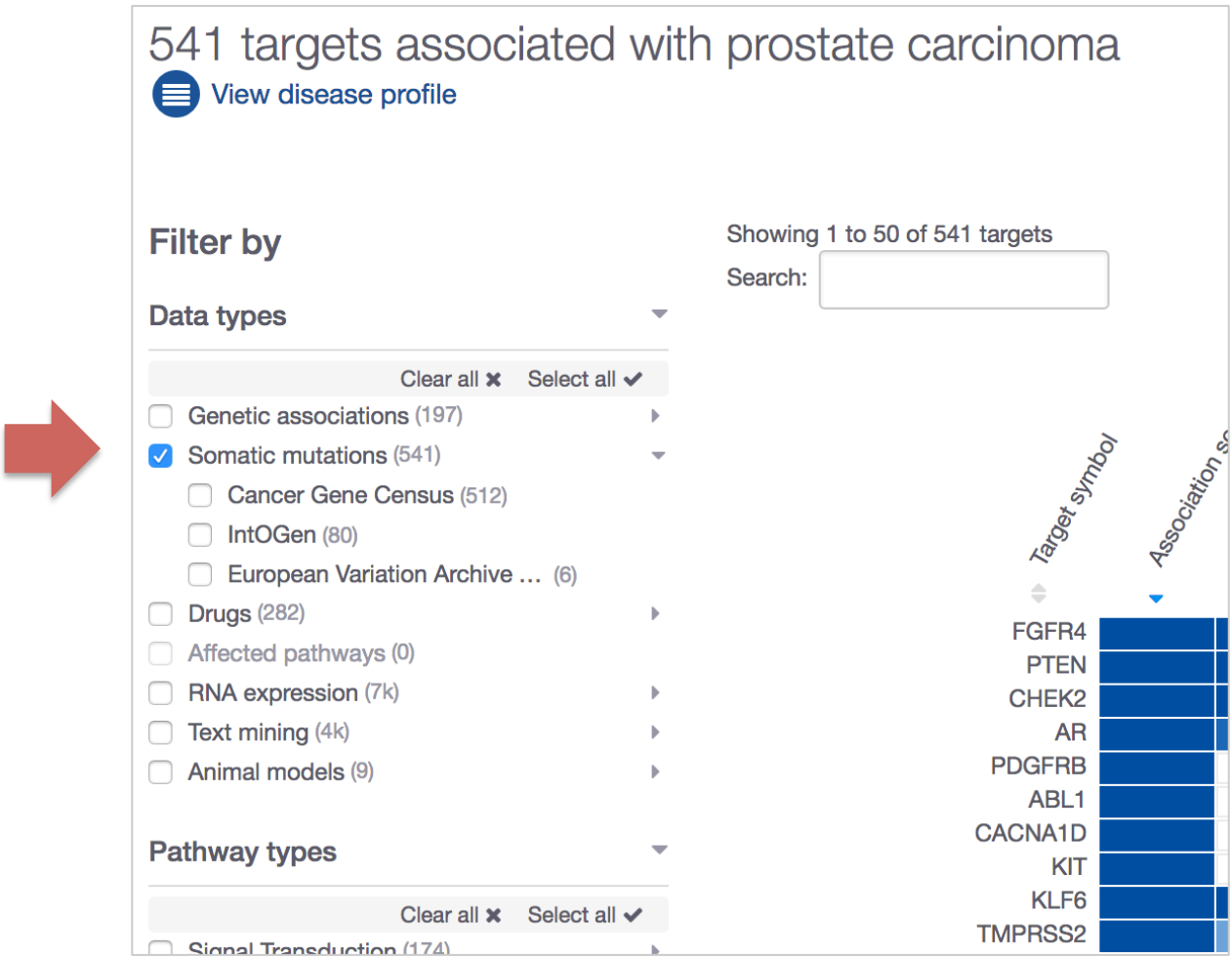
The first 10 rows will show the top 10 targets associated with prostate cancer. These will have the highest score (score of 1): *FGFR4, HOXB13, PTEN*, *CHEK2*, *AR*, *MXI1*, *PDGFRB, ABL1*, *CACNA1D*, and *KIT.*

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).

In a nutshell, the score is computed individually for each piece of evidence (e.g. a drug on phase IV, or a SNP), followed by a score computed for all evidence within a data source (e.g. ChEMBL, GWAS Catalog), then a score for the data type (e.g. Drugs, Genetic associations) and the overall score. The later is shown in the first column of the table above. More details on the scoring can be found in our paper:

nar.oxfordjournals.org/content/early/2016/11/29/nar.gkw1055

b) Apply the Somatic mutations filter to show the targets associated with prostate cancer based on a given data type only:



You can also choose a data source (e.g Cancer Gene Census) to filter the results further.

This filtered list (restricted to somatic mutations only) shows slightly 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. If you were interested in identifying and/or prioritising targets where somatic mutations had been used to support the association, *KLF6* and *TMPRSS2* would be in your top 10 targets.

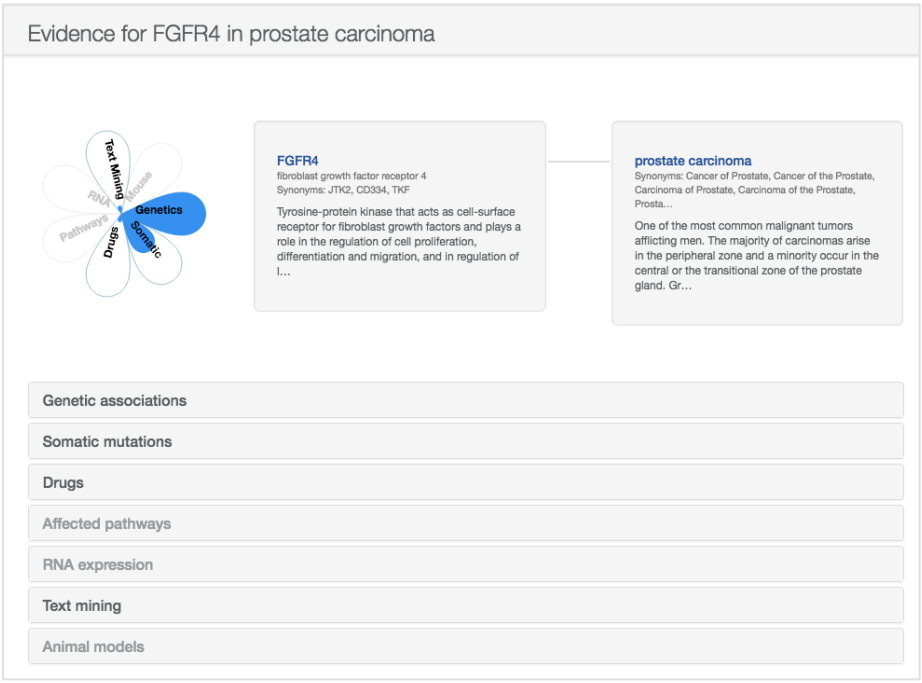
For more details on the data we currently use to associate a gene to a disease can be found below:

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:

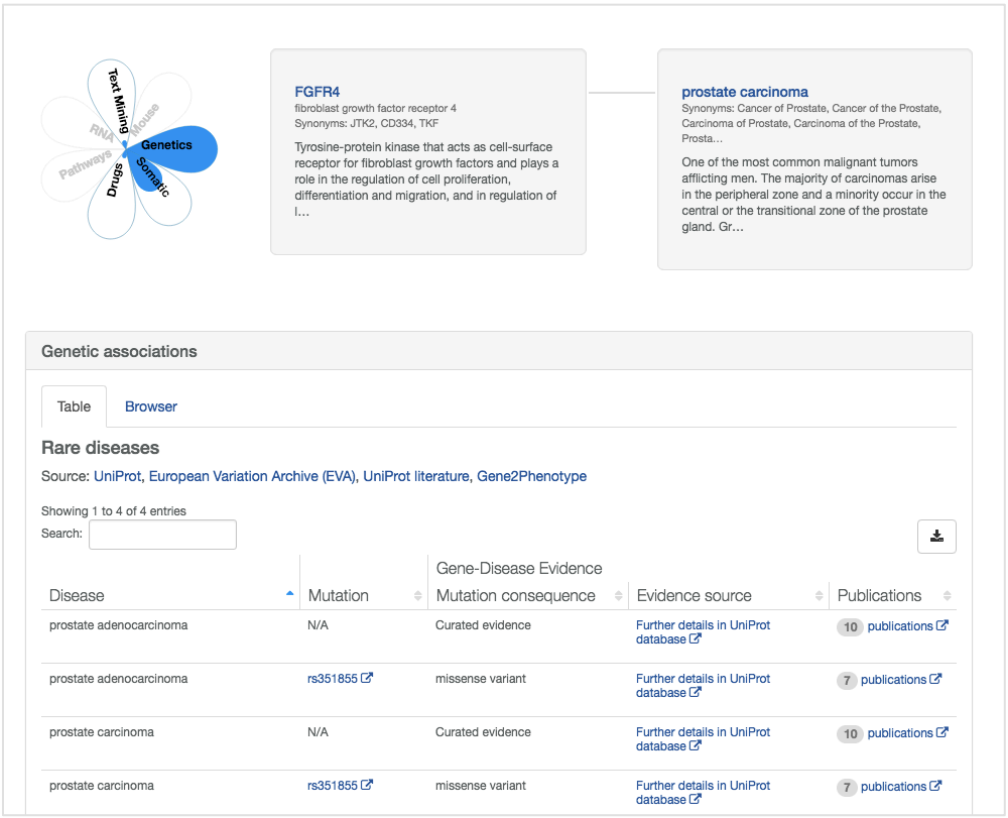


The evidence used to support the association is shown in different tabs (the grey tabs have no data: there is no data for Affected pathways, RNA expression and Animal models to support FGFR4-prostate carcinoma association).

Expand the 'Genetic associations' tab.

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

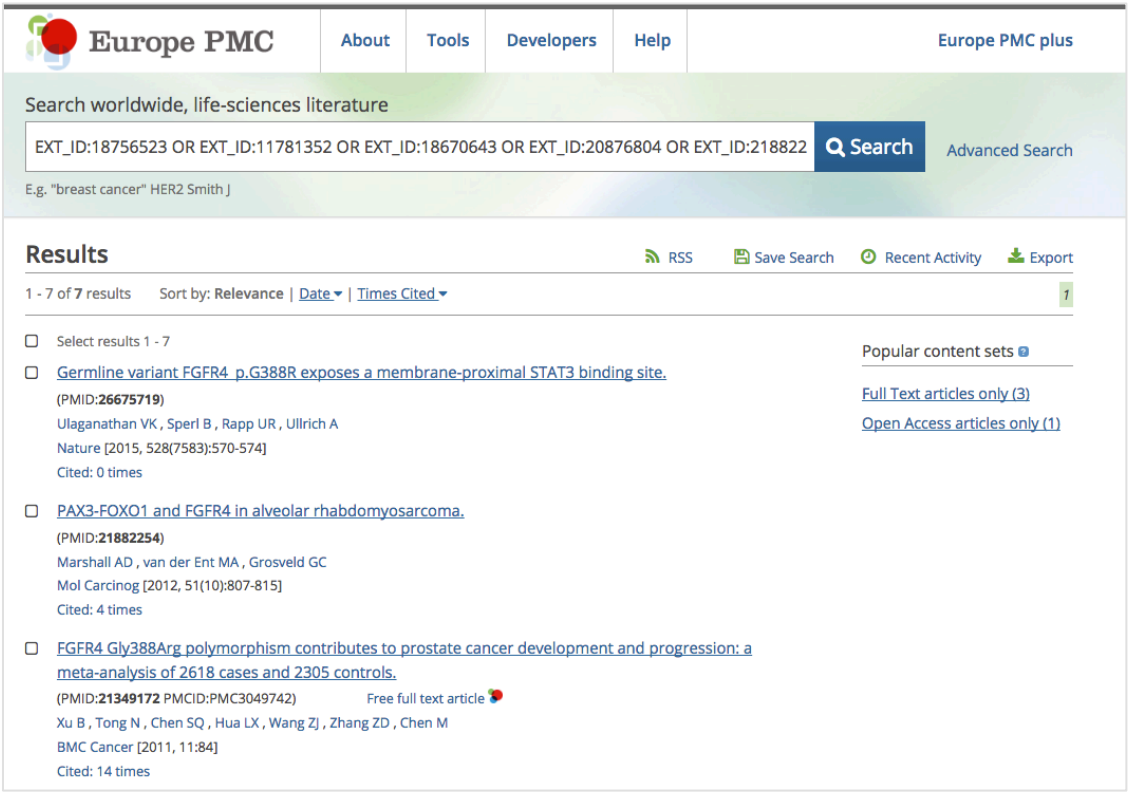
you will land on a page where the tab containing the Genetic association will be already opened:



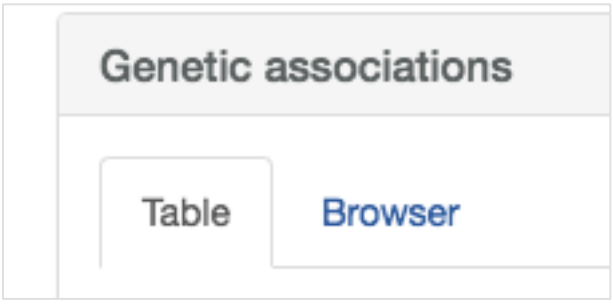
Yes, there is one genetic variant that is known in public databases to be associated with prostate carcinoma. Its ID is rs351855.

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).

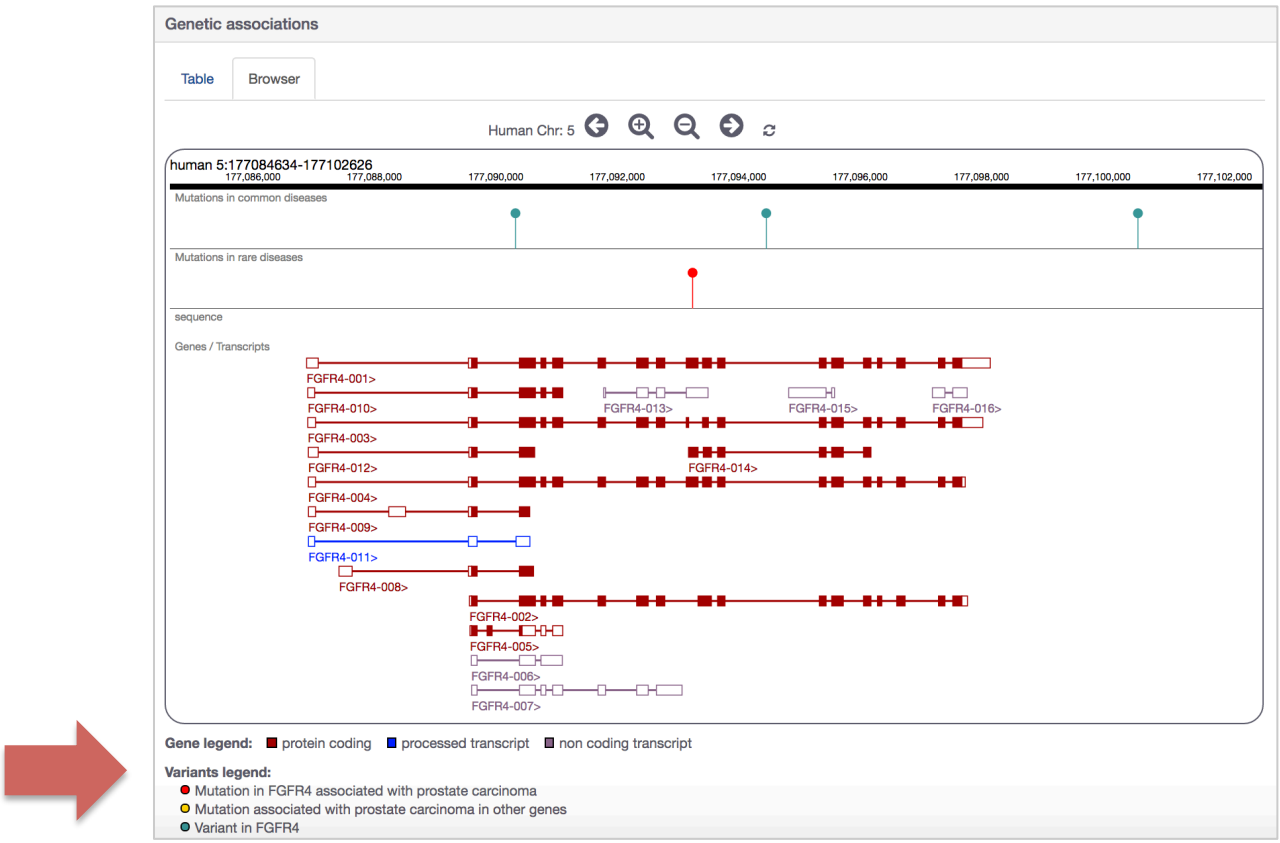
Click on the 7 ‘publications’ link to see the papers supporting the association:



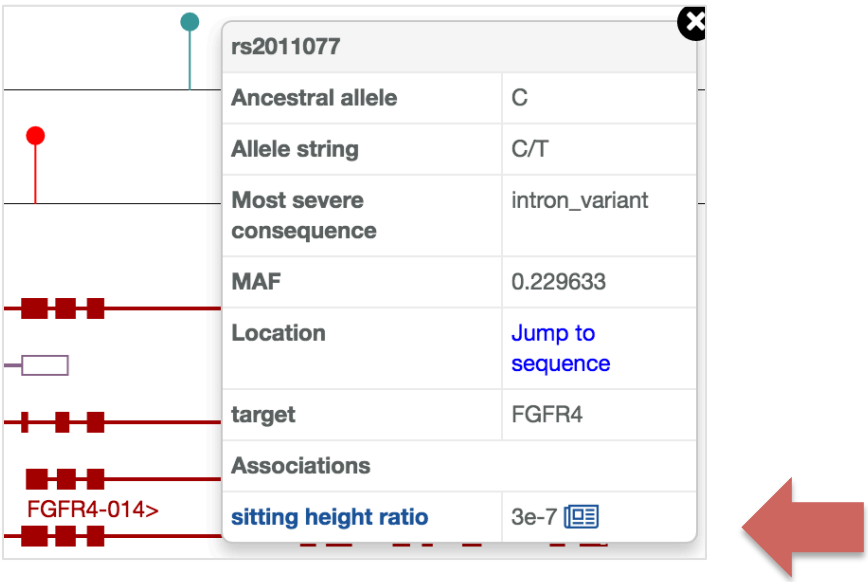
d) The Genetic associations data can also been visualised in a graphical display. Click on the ‘Browser’ link:



You will see the transcripts mapped to that gene and the variants (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.

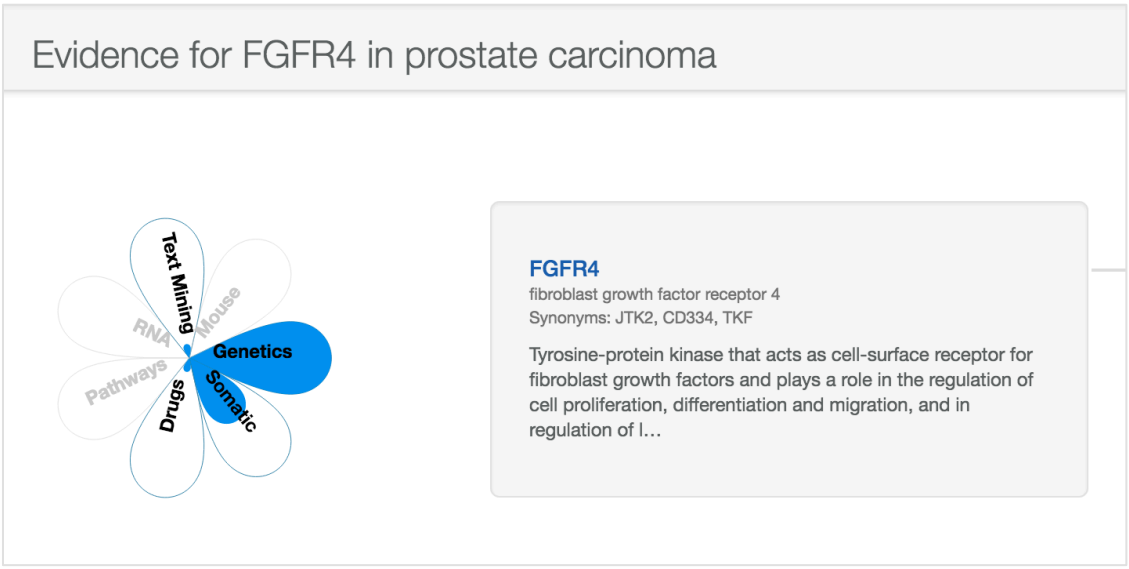


In this genomic region, you can see there are two mutations associated with other traits (i.e. sitting height ratio and body height). You may want to zoom out to view more variants up or downstream of the gene, and then click on the lollipop (the variant) for more details:



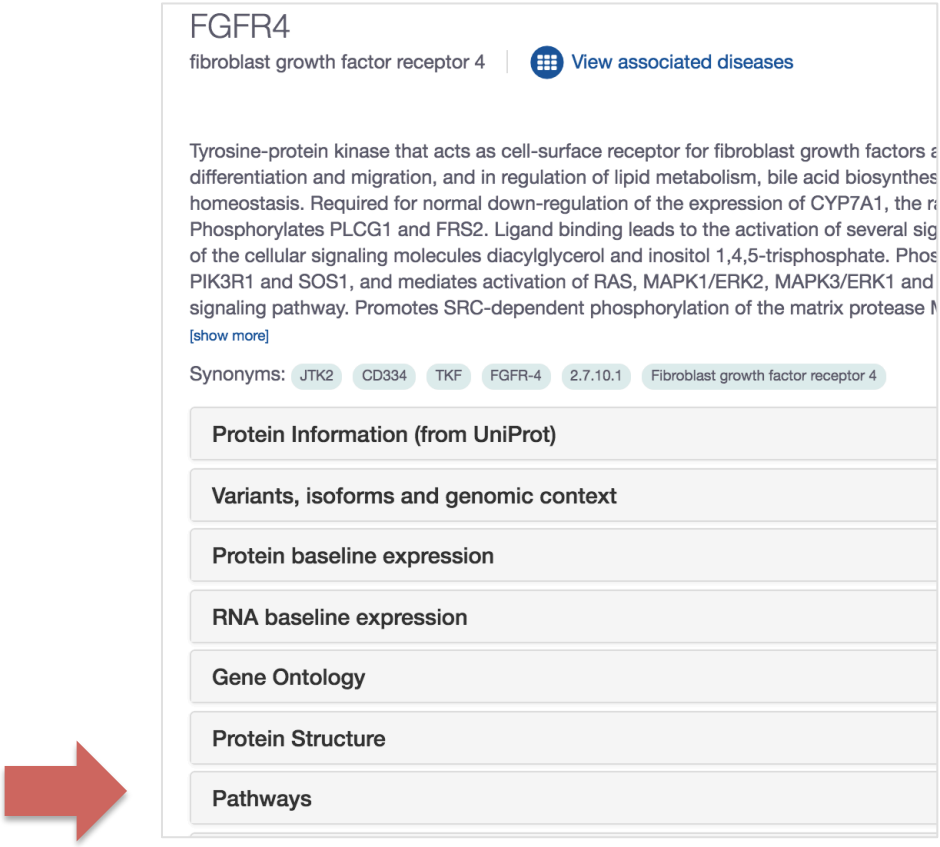
e) Let's now have a look at the target profile page giving us information outside the context of a disease.

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



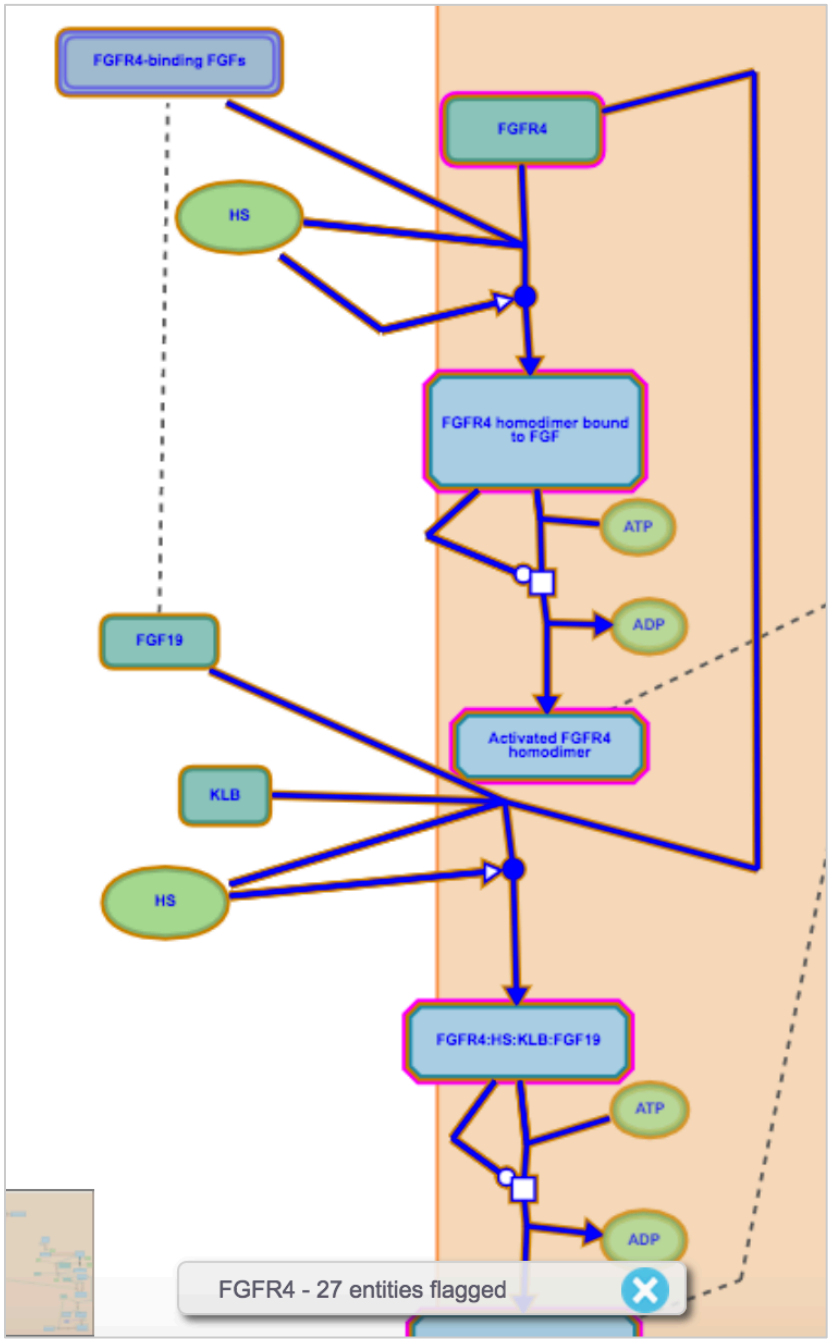
You will end up in a page like this:

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

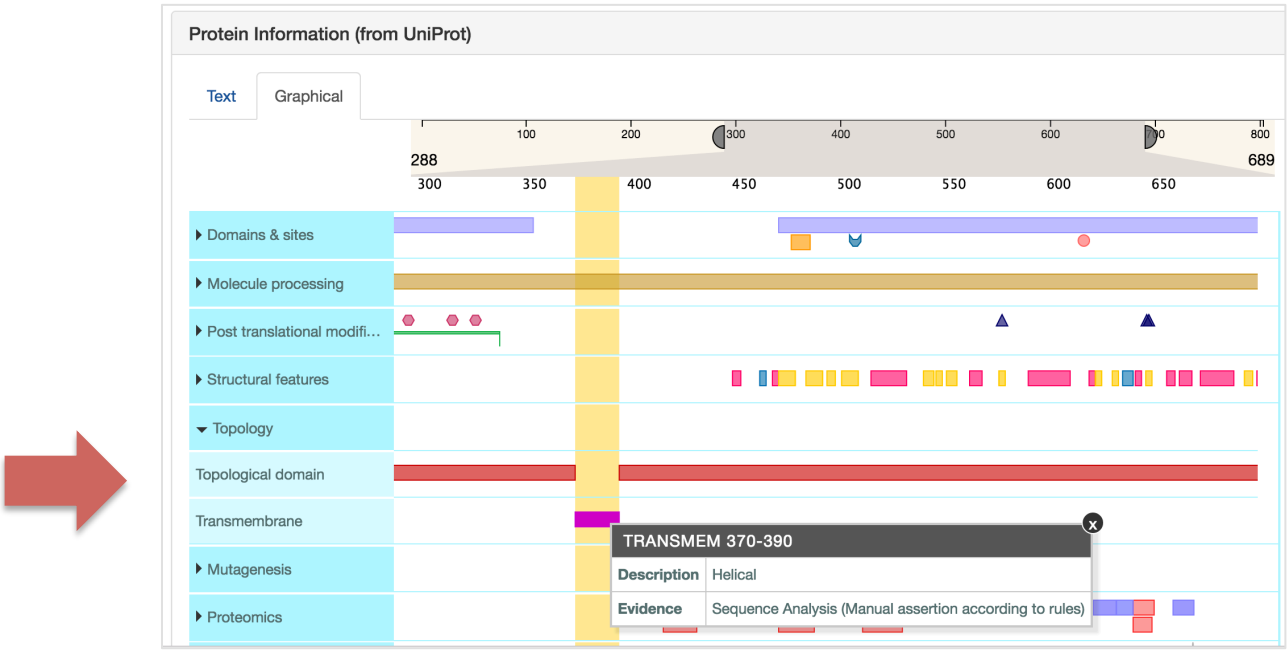


Click on Pathways to find out cellular pathways and biochemical process this gene is involved in e.g. PI3K Cascade, Constitutive Signaling by Aberrant PI3K in Cancer and few others.

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



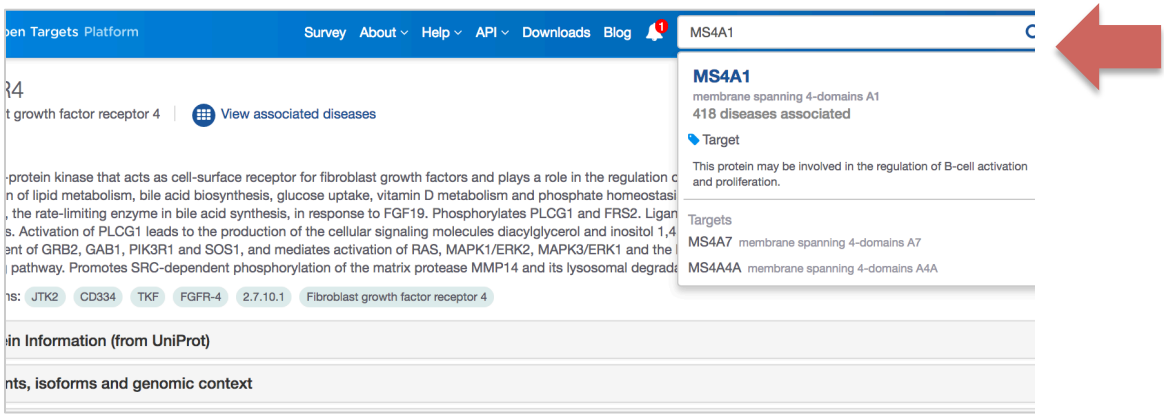
f) Click on the Protein information (from UniProt) tab. Now, click on the Graphical view option, then click on the Topology menu to see the annotated domains: extracellular, transmembrance and intracellular. The transmembrane (TM) domain goes from amino acid 370 to 390. Gene *FGFR4* codes for a receptor,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



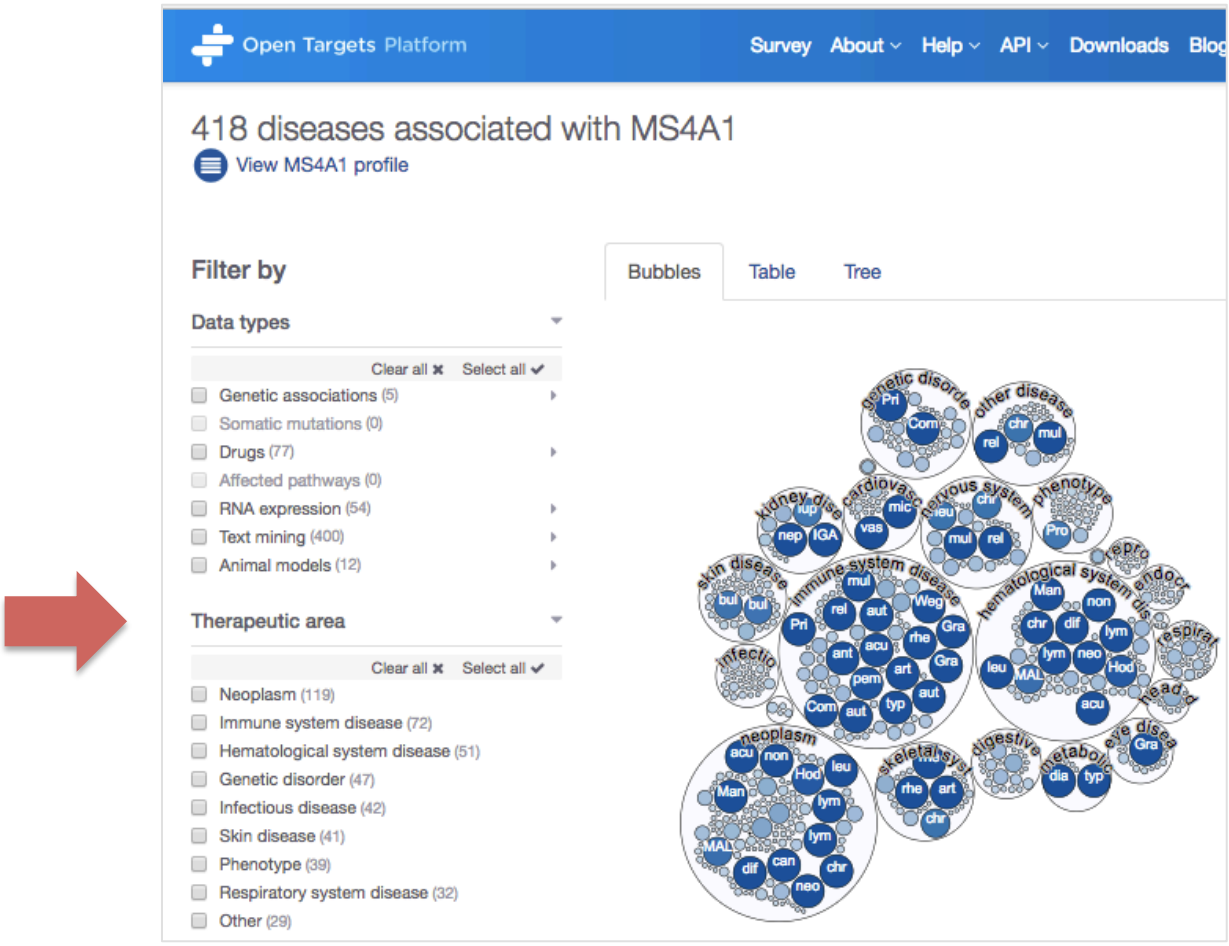
Exercise 2 – *MS4A1* as a possible drug target in the treatment of non-Hodgkin's lymphoma

Search for MS4A1.

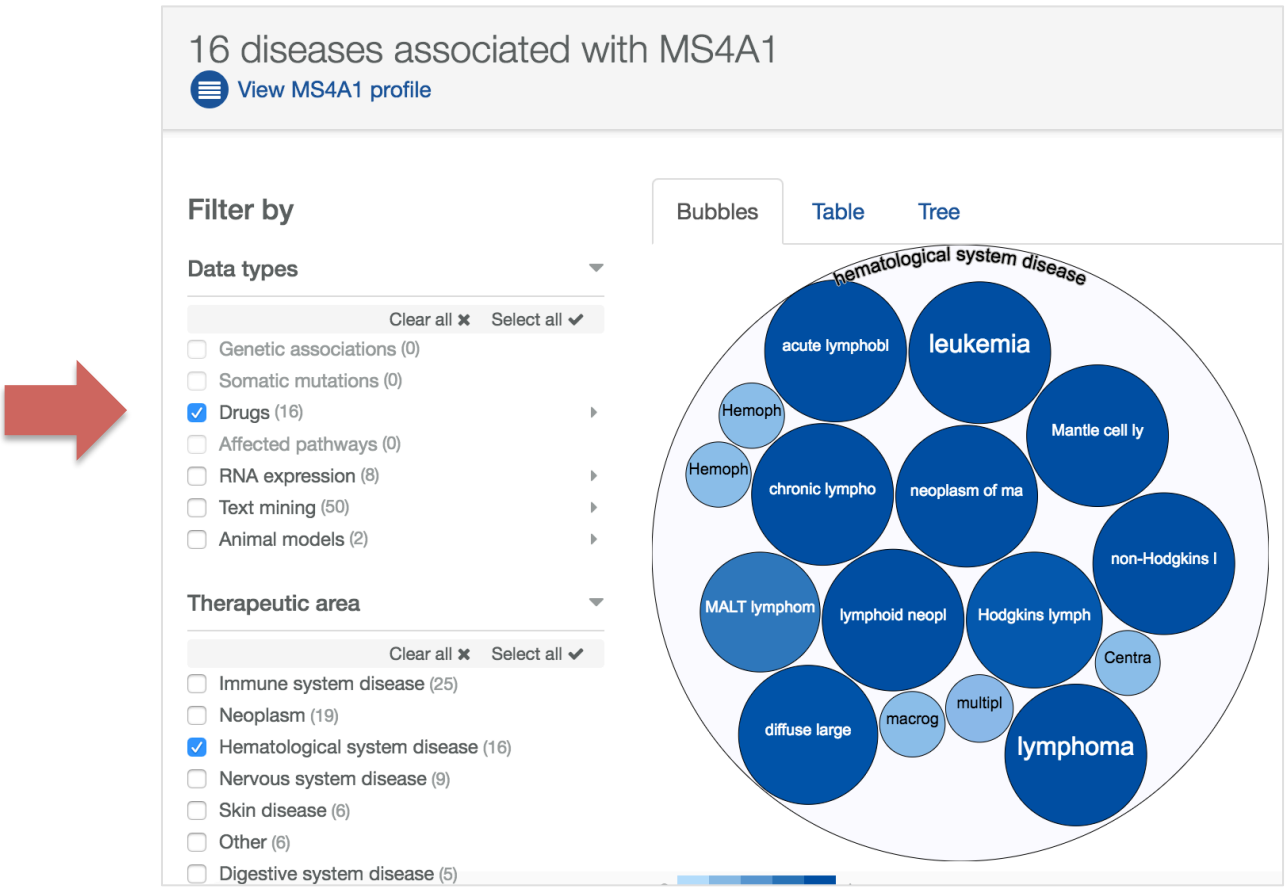
Tip: you do not need to go back to the homepage: you can use the search box at the top right corner of any pages in the Platform:



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:



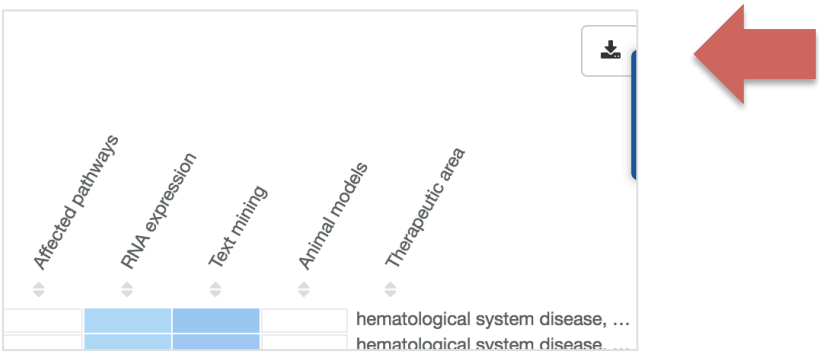
b) The default display for the diseases associated with a target is the Bubbles view.

But you can also view the same information as a Table or Tree. Select the Table view now.

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

Some of the diseases with the (overall) association score of 0.90 or above (for the filters selected above) 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:

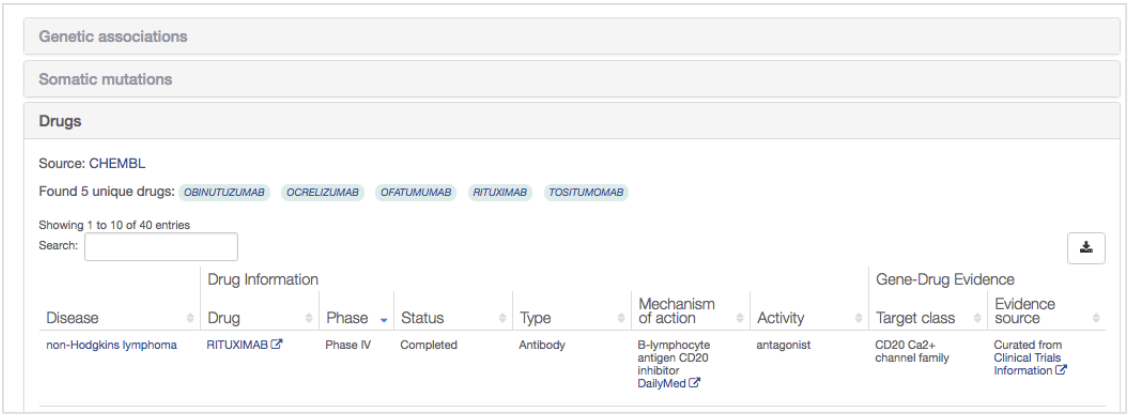
Download the table in CSV.

c) Filter the table with 'non-Hod' for 'non-Hodgkin's lymphoma'. In addition to 'Drugs', evidence from 'Text mining' also supports the association between this disease and *MS4A1*.

Note: the data coming from mining the literature is given a lower weight in our analysis; therefore it will always get a lower score than for example the score for Drugs coming from ChEMBL for example.

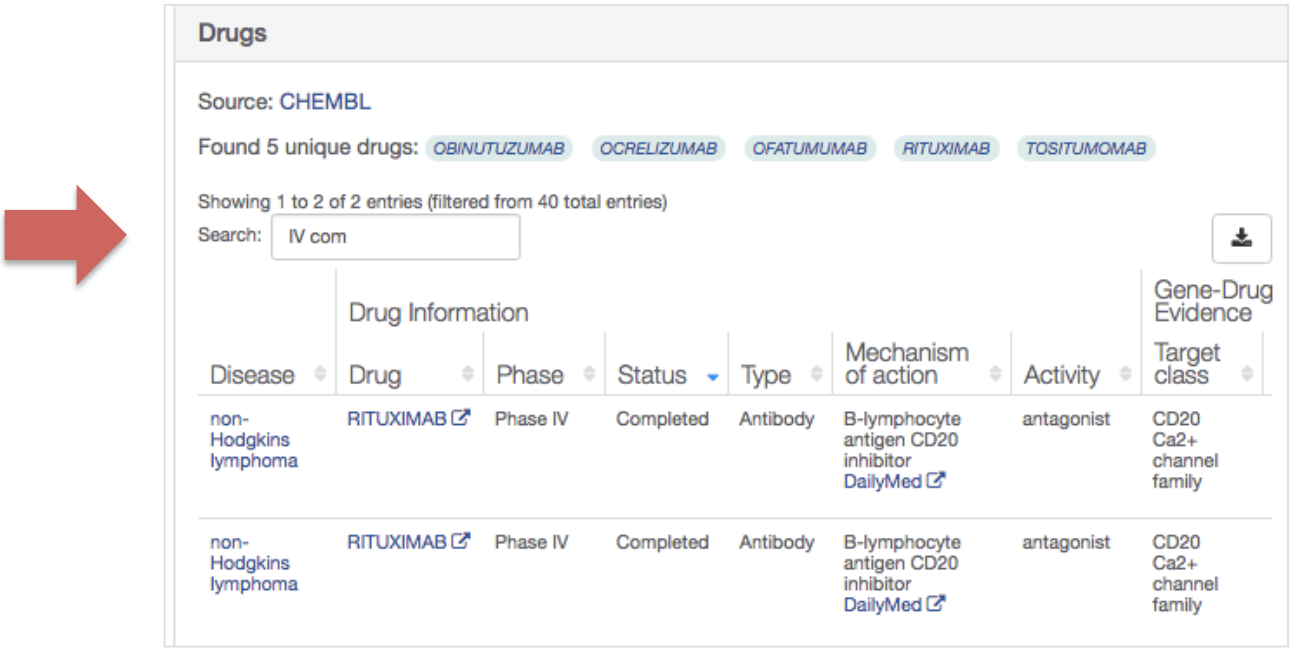
Click on the 'Text mining' cell in the table to see the 778 research articles mined from Pubmed. The text mining pipeline will look for the co-occurrence of the gene name (or its synonym i.e. CD20) and the disease name in the same sentence. This can be in the title, abstract, or other section of the paper with the exception of supplementary tables.

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)

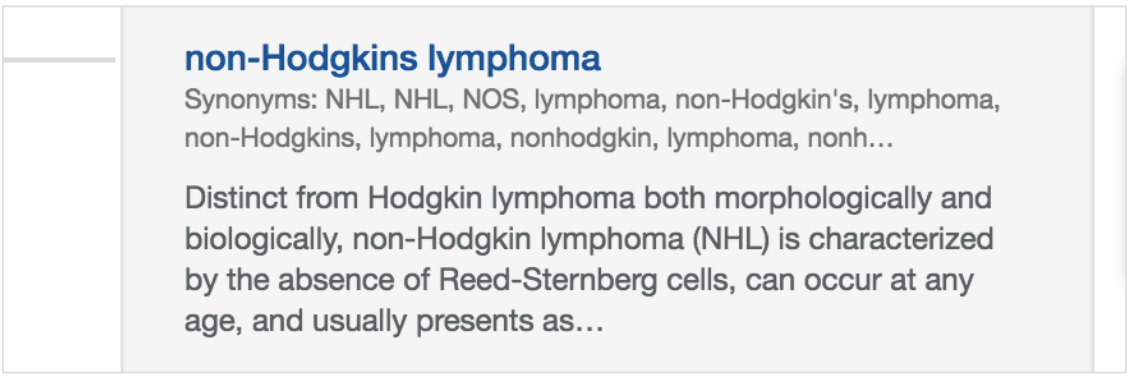
Search: IV com

	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 (apparently) the same information in all the columns?

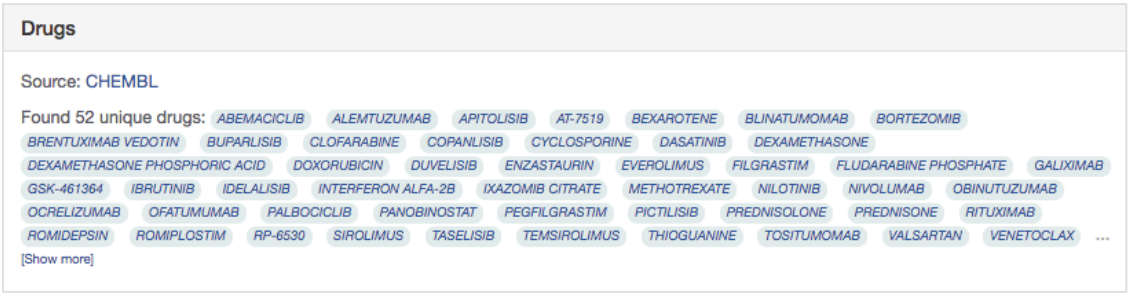
Some rows may look like identical at a first glance. Click on the link in the last column (Evidence source) to find out the drug is actually coming from two different studies, NCT00090038 and NCT00430352. Both are in phase IV, status completed.

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. Then click on the disease name:

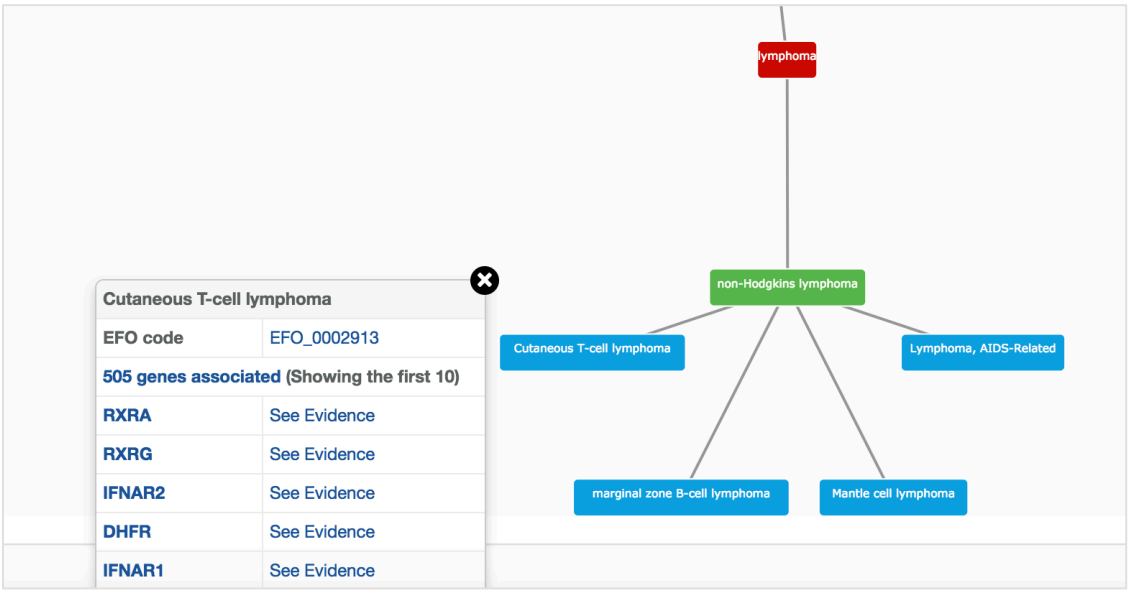


This will take you the disease page with more information on the disease including all drugs marketed for this disease, its ontology and its phenotypes (if available).

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

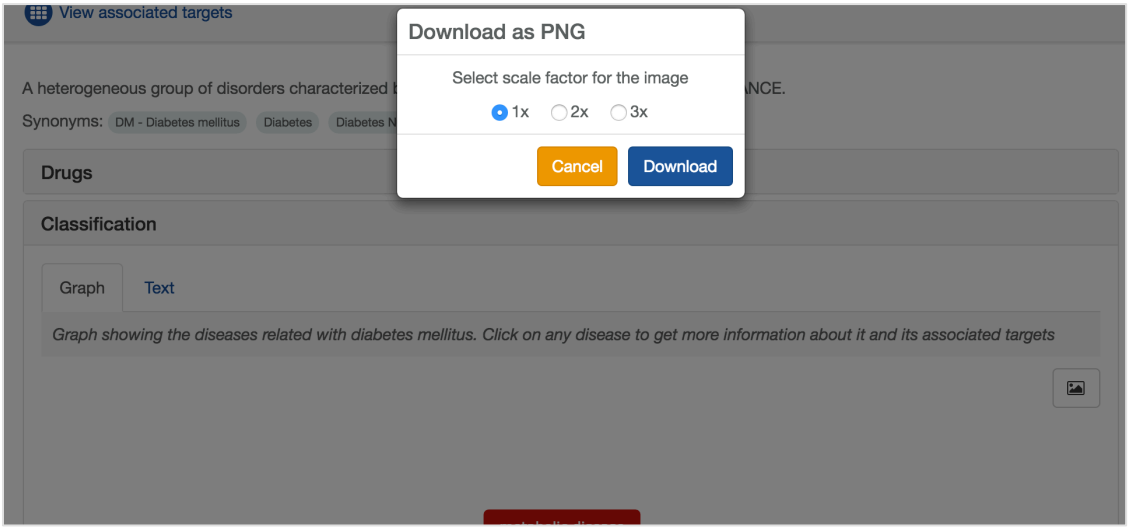


Now, have a look at the 'Classification' tab to find the disease ontology:



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:

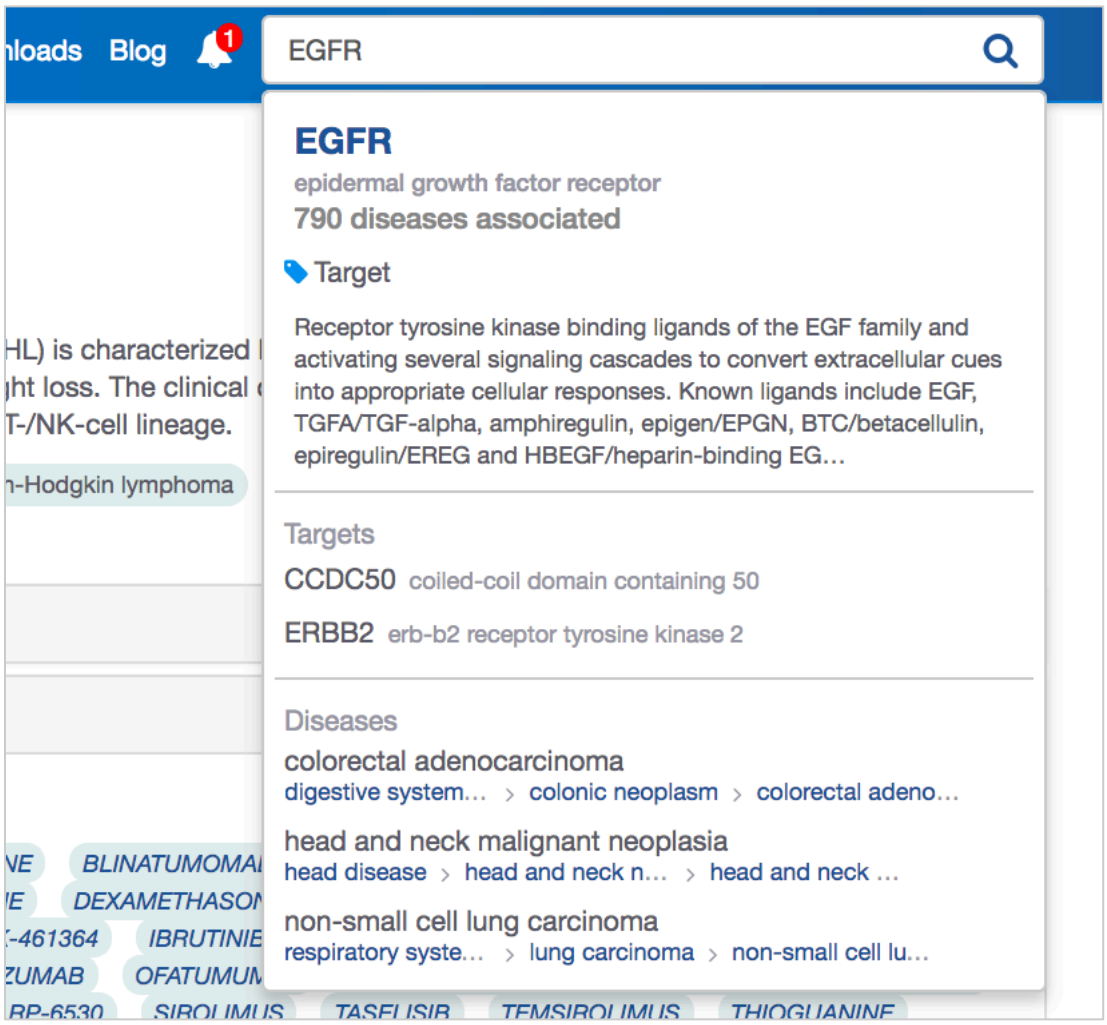


Note some of drugs you have identified in step d) have been investigated in sub-types (children terms) of non-Hodgkin's lymphoma, such as Cutaneous T-cell The diagram makes it easier to see the relationship of the parent disease (non-Hodgkin's lymphoma) and its children diseases (e.g. Cutaneous T-cell lymphoma and Mantle cell lymphoma).

Exercise 3

The *EGFR* gene, a receptor tyrosine kinase

Search for EGFR either from the homepage of the Platform or any page by using the search box at the top right:

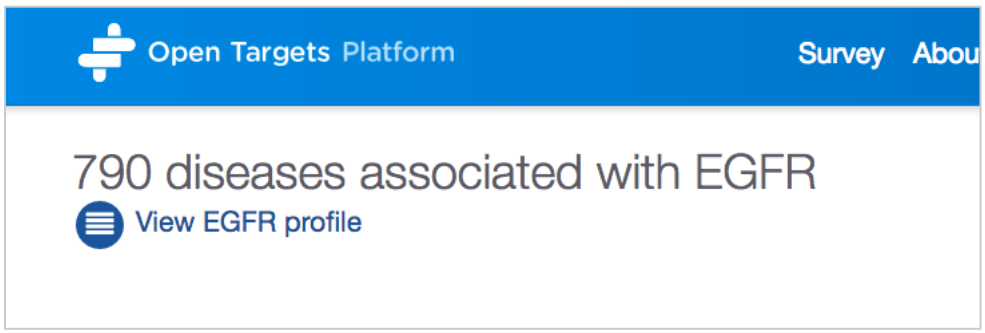


You will then get to this page:

http://targetvalidation.org/target/ENSG00000146648/associations

with all the disease associations with this target.

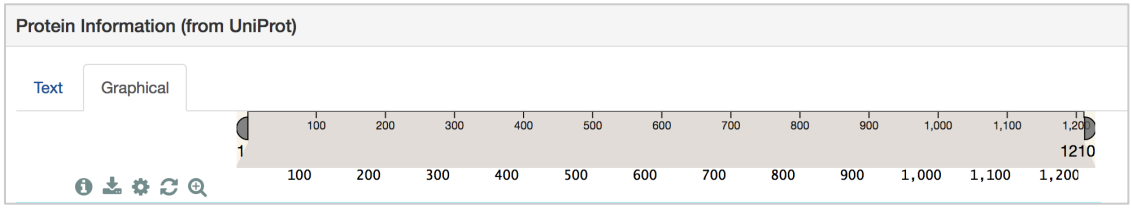
Click on 'View EGFR profile' link:



You are now in the profile page of your target:

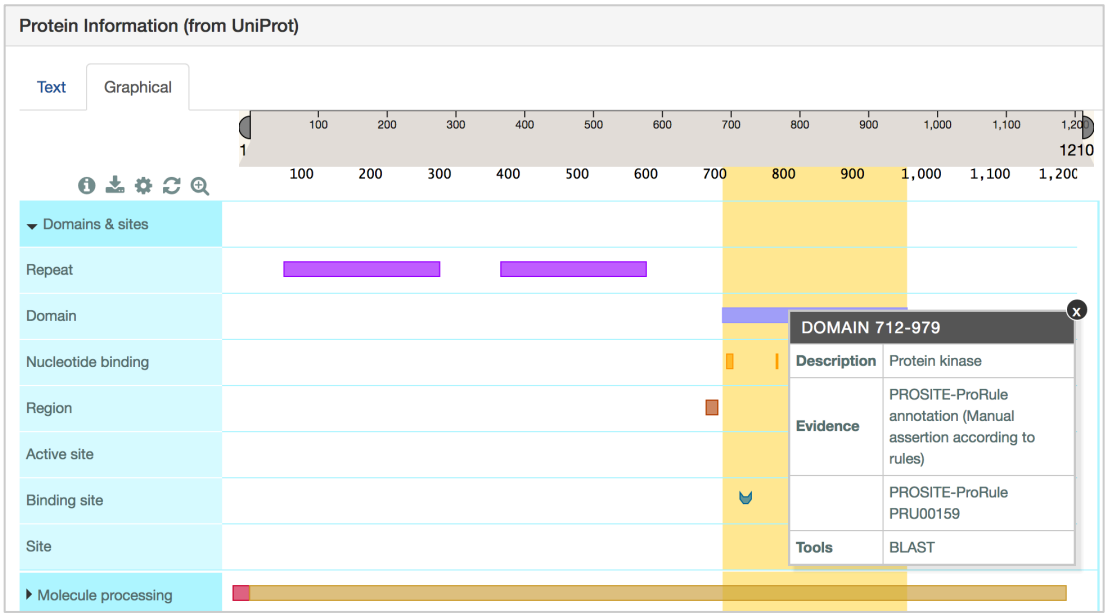
http://targetvalidation.org/target/ENSG00000146648

Expand the "Protein Information (from UniProt)" tab and select the 'Graphical' option:



a) The protein encoded by this gene/target (EGFR, also known as ENSG00000146648) is 1210 amino acid long.

b) Explore the image above and click on the option "Domains & sites":

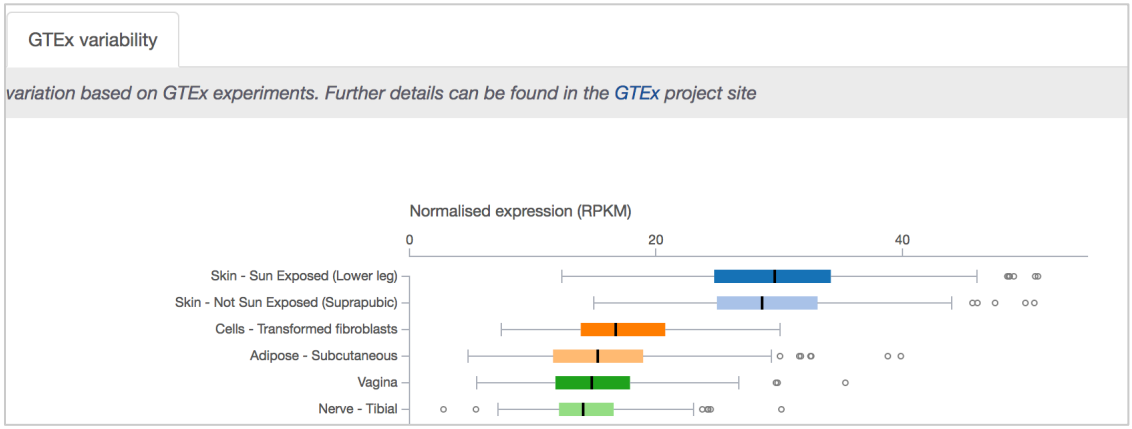


The kinase domain is at position 712-979 in the amino acid sequence. Other sites have been mapped to this protein, e.g. dimerization, phosphorylation and activation sites, ATP binding site and PIK3C2B interaction site.

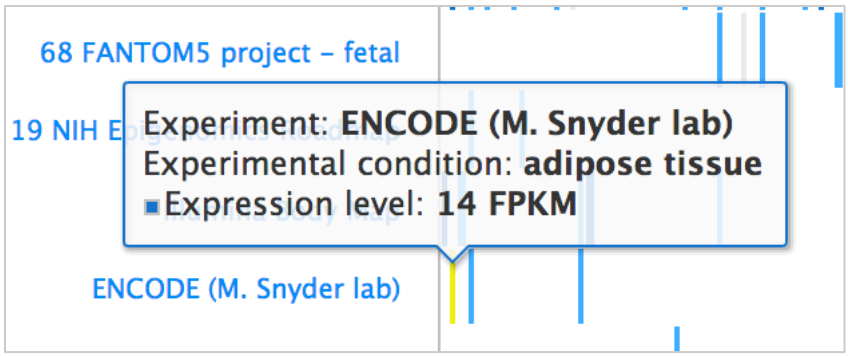
Let's scroll down in the same page and expand the "RNA baseline expression" tab:

c) The tissues with the highest RNA baseline expression are:

1) skin from the GTEx project:



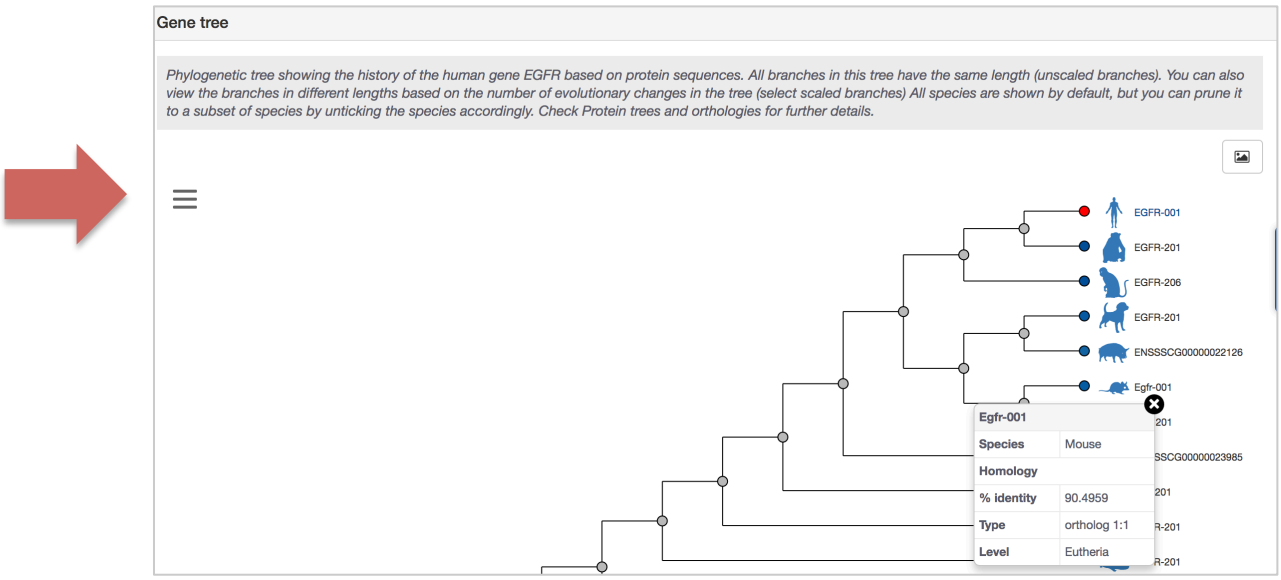
2) adipose tissue from the ENCODE project available in Expression Atlas:



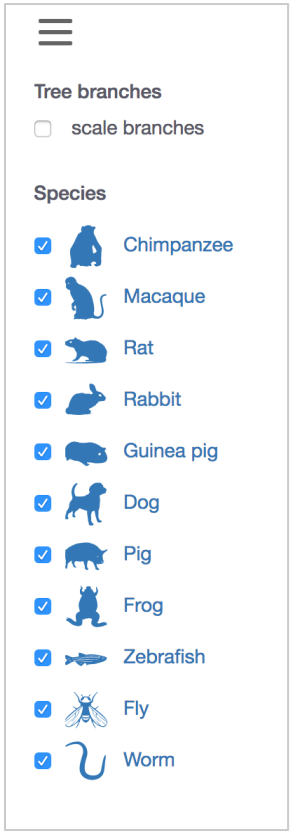
d) Expand the "Gene Ontology" tab to find out some of the molecular functions the *EGFR* protein e.g. actin filament binding, transmembrane receptor protein tyrosine kinase activity, MAP kinase kinase kinase activity, to name a few.

e) Yes, you can use the Open Targets Platform to find out if there is a mouse orthologue for this human gene.

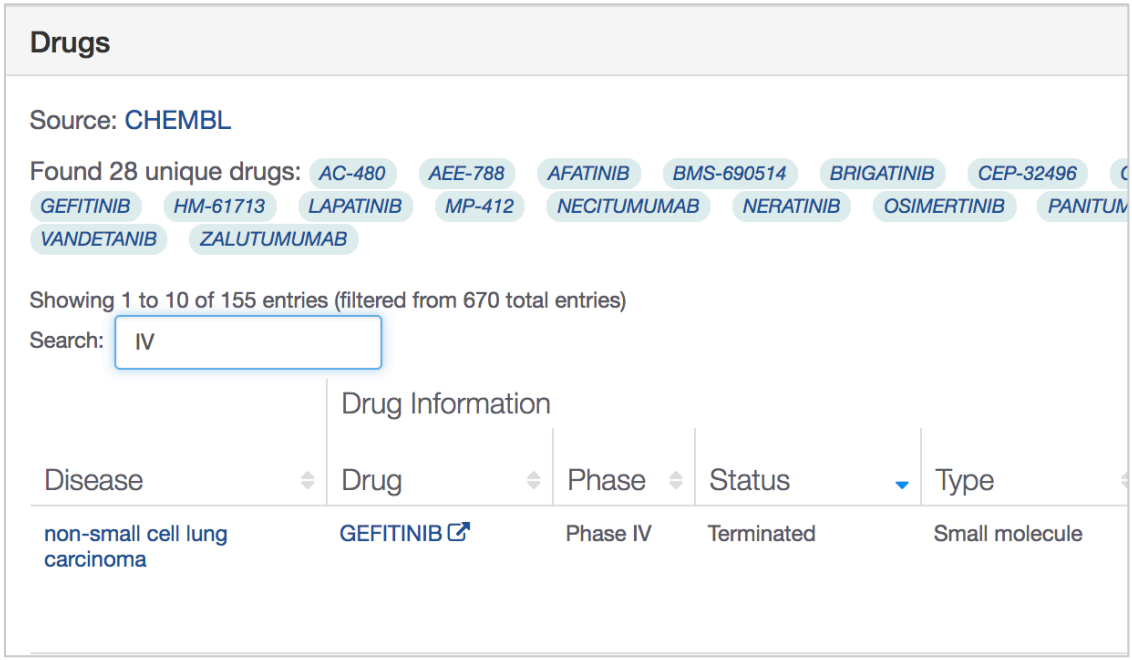
Expand the "Gene tree" tab to see a phylogenetic tree for this gene:



On the menu icon at the left, you can select which species to display in the tree. You can also change from unscaled branches (the default choice: all branches with the same length) to scaled branches (where the branches will be displayed in different lengths based on the number of evolutionary changes in the tree):



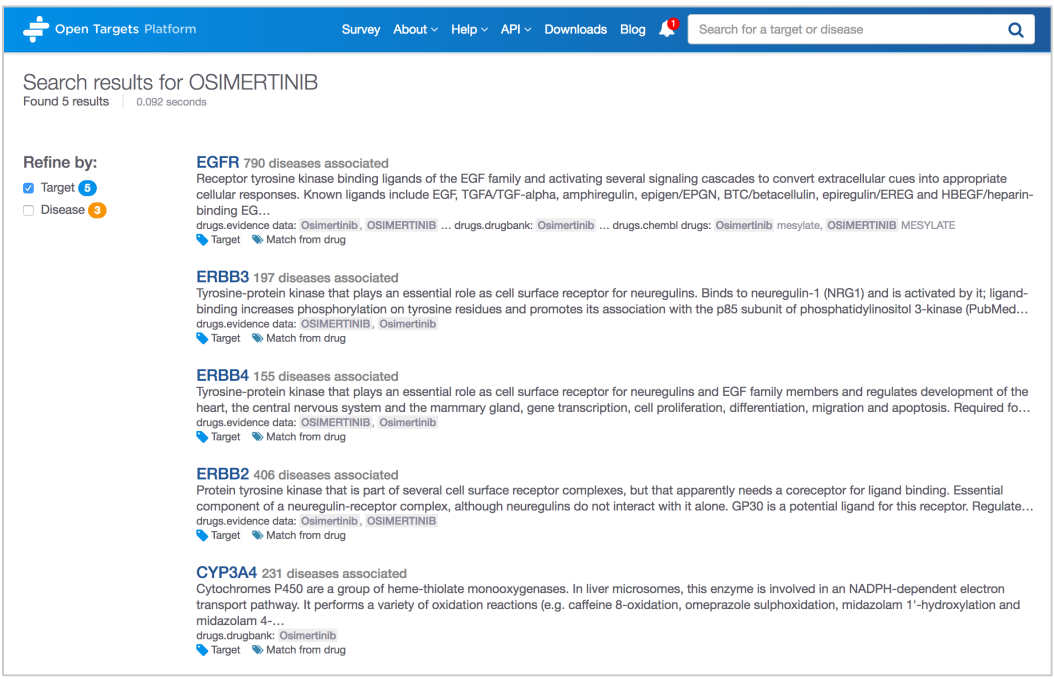
f) Still on the same page, expand the "Drugs" tab to find out that GEFITINIB is the drug in clinical trial phase IV, status terminated, which is targeting this gene in studies with non-small cell lung carcinoma patients:



The last column in the table provides links to the original study from the NIH Clinical Trials site:

https://clinicaltrials.gov/search?id=%22NCT00536107%22

g) Search for OSIMERTINIB to find out that this drug does target other kinases than *EGFR*, i.e. *ERBB3, ERBB2* and *ERBB4.* **Note:** this drug (and perhaps many others) can match CYP3A4 (also known as P450). p450 does metabolise the OSIMERTINIB but it is not the target for that drug. Check the source of the match (drugs.evidence data for the different kinases *versus drugs.drugbank* for P450).


Open Targets Platform
Survey
About
Help
API
Downloads
Blog
Search for a target or disease
Search results for OSIMERTINIB
Found 5 results
0.092 seconds
Refine by:
Target 5
Disease 3
EGFR 790 diseases associated
Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades to convert extracellular cues into appropriate cellular responses. Known ligands include EGF, TGFA/TGF-alpha, amphiregulin, epigen/EPGN, BTC/betacellulin, epiregulin/EREG and HBEGF/heparin-binding EG…
drugs.evidence data: Osimertinib, OSIMERTINIB … drugs.drugbank: Osimertinib … drugs.chembl drugs: Osimertinib mesylate, OSIMERTINIB MESYLATE
Target
Match from drug
ERBB3 197 diseases associated
Tyrosine-protein kinase that plays an essential role as cell surface receptor for neuregulins. Binds to neuregulin-1 (NRG1) and is activated by it; ligand-binding increases phosphorylation on tyrosine residues and promotes its association with the p85 subunit of phosphatidylinositol 3-kinase (PubMed…
drugs.evidence data: OSIMERTINIB, Osimertinib
Target
Match from drug
ERBB4 155 diseases associated
Tyrosine-protein kinase that plays an essential role as cell surface receptor for neuregulins and EGF family members and regulates development of the heart, the central nervous system and the mammary gland, gene transcription, cell proliferation, differentiation, migration and apoptosis. Required fo…
drugs.evidence data: OSIMERTINIB, Osimertinib
Target
Match from drug
ERBB2 406 diseases associated
Protein tyrosine kinase that is part of several cell surface receptor complexes, but that apparently needs a coreceptor for ligand binding. Essential component of a neuregulin-receptor complex, although neuregulins do not interact with it alone. GP30 is a potential ligand for this receptor. Regulate…
drugs.evidence data: Osimertinib, OSIMERTINIB
Target
Match from drug
CYP3A4 231 diseases associated
Cytochromes P450 are a group of heme-thiolate monooxygenases. In liver microsomes, this enzyme is involved in an NADPH-dependent electron transport pathway. It performs a variety of oxidation reactions (e.g. caffeine 8-oxidation, omeprazole sulphoxidation, midazolam 1'-hydroxylation and midazolam 4-…
drugs.drugbank: Osimertinib
Target
Match from drug