

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*:

The screenshot shows the Open Targets Platform interface. A search bar at the top contains the text "prostate". Below the search bar, the heading "Find new targets for drug discovery" is displayed. Underneath, a section for "prostate carcinoma" is shown, stating "9116 targets associated". A "Disease" section describes prostate carcinoma as one of the most common malignant tumors in men, noting its occurrence in the peripheral zone and central/transitional zone of the prostate gland. Below this, a "Targets" section lists NKX3-1, SPDEF, and KLK3. Further down, a "Diseases" section lists prostate adenocarcinoma, metastatic prostate cancer, and urogenital neoplasm.

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

Open Targets Platform Survey About Help API Downloads Blog Search for a target or disease

9116 targets associated with prostate carcinoma
[View disease profile](#)

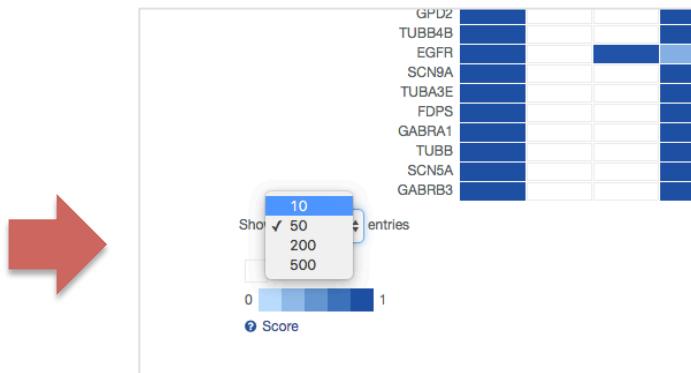
Filter by

Data types

Showing 1 to 50 of 9,116 targets
Search:

Target symbol	Association score	Genetic associations	Somatic mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Target name
FGFR4									fibroblast growth factor re...
HOXB13									homeobox B13
PTEN									phosphatase and tensin h...
CHEK2									checkpoint kinase 2
AR									androgen receptor
MXI1									MAX interactor 1, dimeriz...
PDGFRB									platelet derived growth fa...
ABL1									ABL proto-oncogene 1, n...
CACNA1D									calcium voltage-gated ch...
KIT									KIT proto-oncogene rece...
KLF6									Kruppel like factor 6

Scroll down and select to see 10 entries (rows) only in the result table:

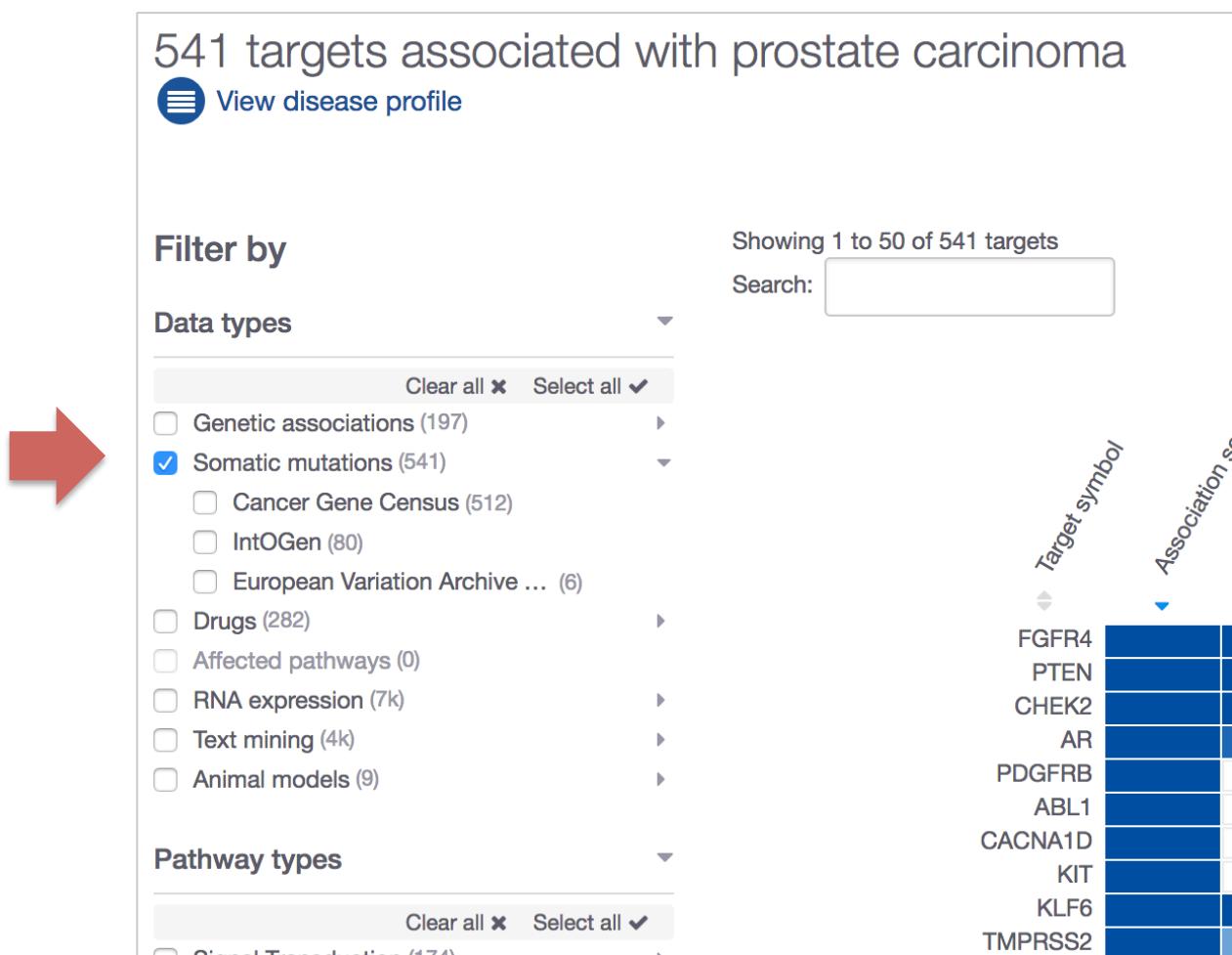


The first 10 rows will show the top 10 targets associated with prostate cancer. These will have the highest score (score of 1): *FGFR4*, *HOXB13*, *PTEN*, *CHEK2*, *AR*, *MXI1*, *PDGFRB*, *ABL1*, *CACNA1D*, and *KIT*.

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:

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



541 targets associated with prostate carcinoma

[View disease profile](#)

Filter by

Data types

Showing 1 to 50 of 541 targets

Search:

Genetic associations (197) ▾

Somatic mutations (541) ▾

- Cancer Gene Census (512)
- IntOGen (80)
- European Variation Archive ... (6)

Drugs (282) ▾

Affected pathways (0)

RNA expression (7k) ▾

Text mining (4k) ▾

Animal models (9) ▾

Pathway types

Clear all × Select all ✓

Signal Transduction (174) ▾

Target symbol	Association score
FGFR4	High
PTEN	Medium
CHEK2	Medium
AR	Medium
PDGFRB	Medium
ABL1	Medium
CACNA1D	Medium
KIT	Medium
KLF6	Medium
TMPRSS2	Medium

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 screenshot shows a search result for "Evidence for FGFR4 in prostate carcinoma". On the left, there is a circular navigation menu with tabs: Genetics, Somatic mutations, Drugs, Affected pathways, RNA expression, Text mining, and Animal models. The "Genetics" tab is highlighted. In the center, there are two main boxes. The left box contains information about the gene "FGFR4": its full name (fibroblast growth factor receptor 4), synonyms (JTK2, CD334, TKF), and a brief description of its function as a cell-surface receptor. The right box contains information about "prostate carcinoma": its synonyms (Cancer of Prostate, Cancer of the Prostate, Carcinoma of Prostate, Carcinoma of the Prostate, Prosta...), a brief description of it being one of the most common malignant tumors afflicting men, and a truncated text entry "...".

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:

Disease	Mutation	Gene-Disease Evidence	Mutation consequence	Evidence source	Publications
prostate adenocarcinoma	N/A	Curated evidence		Further details in UniProt database	10 publications
prostate adenocarcinoma	rs351855		missense variant	Further details in UniProt database	7 publications
prostate carcinoma	N/A	Curated evidence		Further details in UniProt database	10 publications
prostate carcinoma	rs351855		missense variant	Further details in UniProt database	7 publications

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).

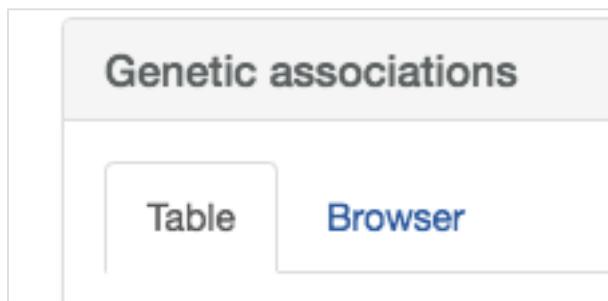
Click on the 7 ‘publications’ link to see the papers supporting the association:

The screenshot shows the Europe PMC search interface. The search bar contains the query "EXT_ID:18756523 OR EXT_ID:11781352 OR EXT_ID:18670643 OR EXT_ID:20876804 OR EXT_ID:218822". Below the search bar is a placeholder text "E.g. "breast cancer" HER2 Smith J". The results section displays three publications:

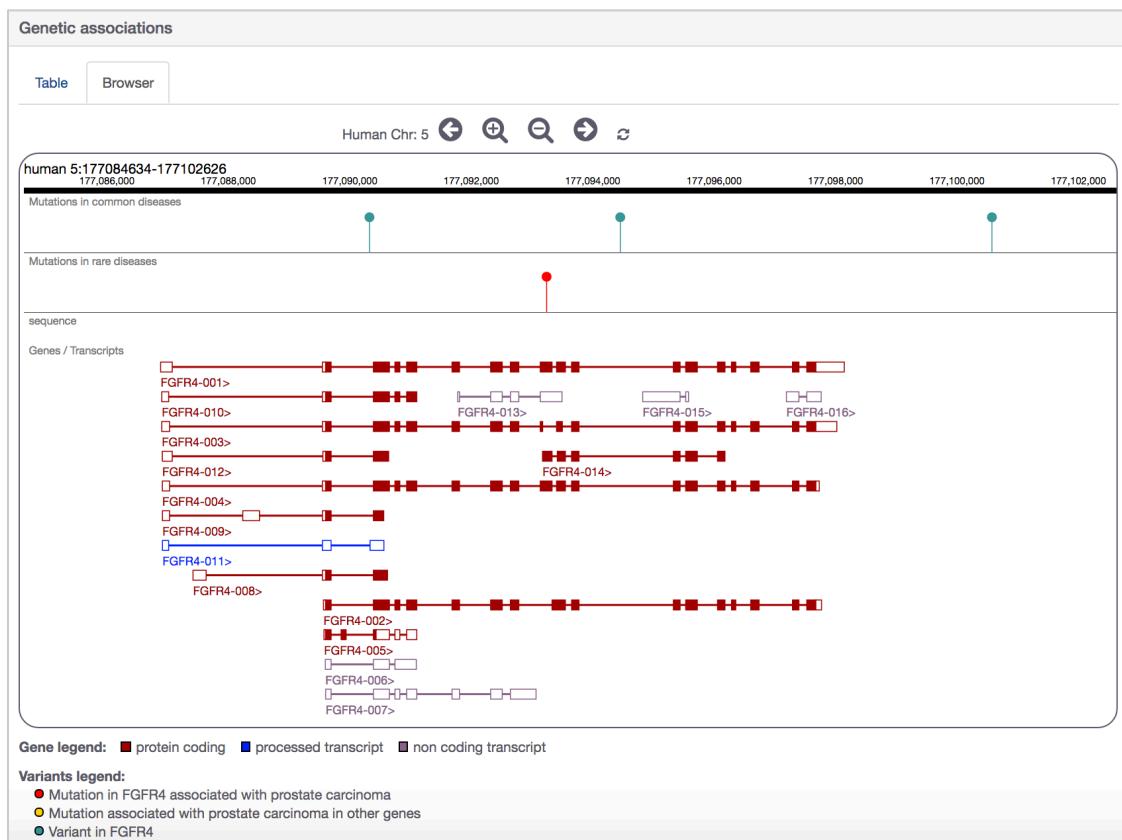
- [Germline variant FGFR4 p.G388R exposes a membrane-proximal STAT3 binding site.](#) (PMID:26675719)
Ulaganathan VK, Sperl B, Rapp UR, Ullrich A
Nature [2015, 528(7583):570-574]
Cited: 0 times
- [PAX3-FOXO1 and FGFR4 in alveolar rhabdomyosarcoma.](#) (PMID:21882254)
Marshall AD, van der Ent MA, Grosfeld GC
Mol Carcinog [2012, 51(10):807-815]
Cited: 4 times
- [FGFR4 Gly388Arg polymorphism contributes to prostate cancer development and progression: a meta-analysis of 2618 cases and 2305 controls.](#) (PMID:21349172 PMCID:PMC3049742)
Xu B, Tong N, Chen SQ, Hua LX, Wang ZJ, Zhang ZD, Chen M
BMC Cancer [2011, 11:84]
Cited: 14 times

On the right side of the results page, there are links for "Popular content sets" (with 3 items), "Full Text articles only (3)", and "Open Access articles only (1)".

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:

Evidence for FGFR4 in prostate carcinoma

FGFR4
fibroblast growth factor receptor 4
Synonyms: JTK2, CD334, TKF
Tyrosine-protein kinase that acts as cell-surface receptor for fibroblast growth factors and plays a role in the regulation of cell proliferation, differentiation and migration, and in regulation of l...

You will end up in a page like this:

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

FGFR4
fibroblast growth factor receptor 4 | [View associated diseases](#)

Tyrosine-protein kinase that acts as cell-surface receptor for fibroblast growth factors and plays a role in cell proliferation, differentiation and migration, and in regulation of lipid metabolism, bile acid biosynthesis and homeostasis. Required for normal down-regulation of the expression of CYP7A1, the rate-limiting enzyme in the synthesis of bile acids. Phosphorylates PLCG1 and FRS2. Ligand binding leads to the activation of several signaling molecules diacylglycerol and inositol 1,4,5-trisphosphate. Phosphorylates PIK3R1 and SOS1, and mediates activation of RAS, MAPK1/ERK2, MAPK3/ERK1 and PI3K-Akt signaling pathway. Promotes SRC-dependent phosphorylation of the matrix protease Mmp9. [show more]

Synonyms: [JTK2](#) [CD334](#) [TKF](#) [FGFR-4](#) [2.7.10.1](#) [Fibroblast growth factor receptor 4](#)

Protein Information (from UniProt)

[Variants, isoforms and genomic context](#)

[Protein baseline expression](#)

[RNA baseline expression](#)

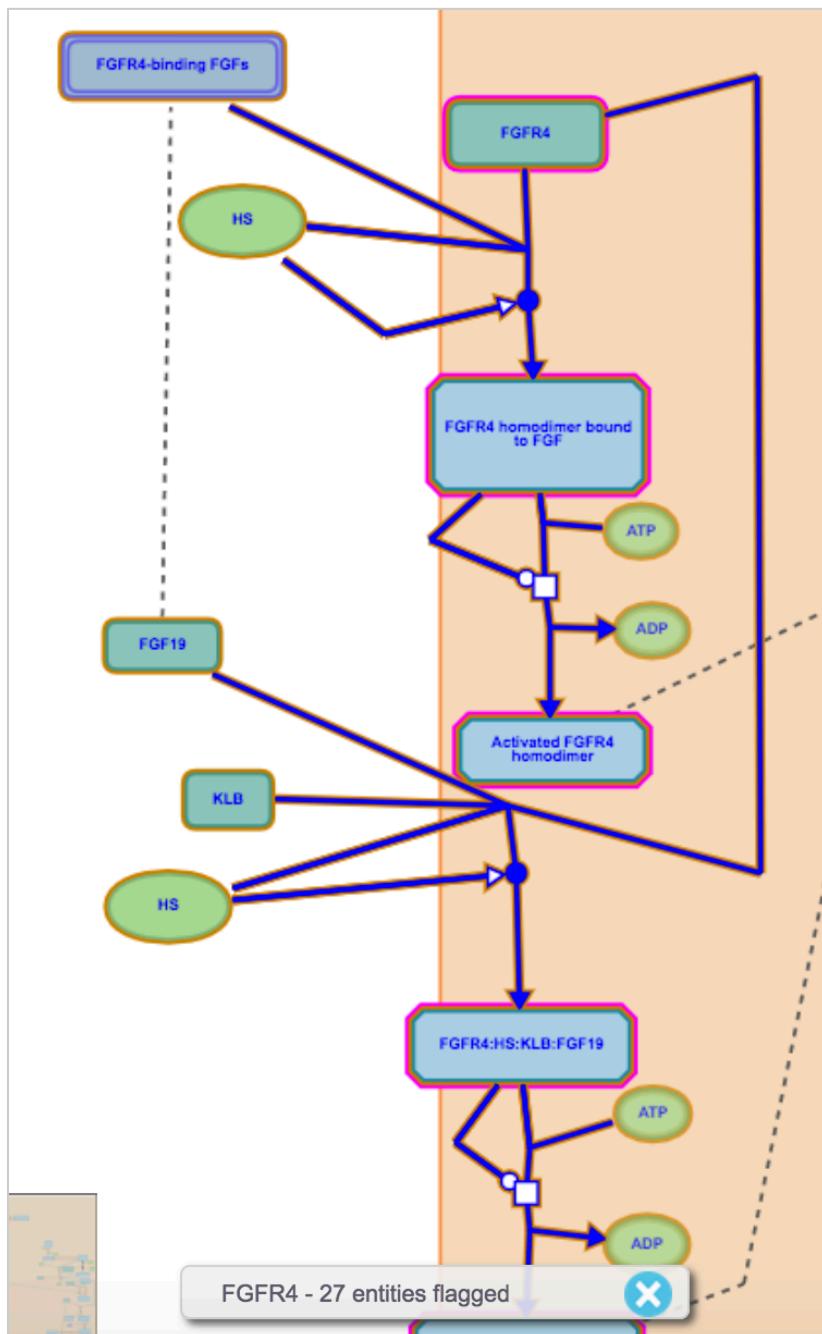
[Gene Ontology](#)

[Protein Structure](#)

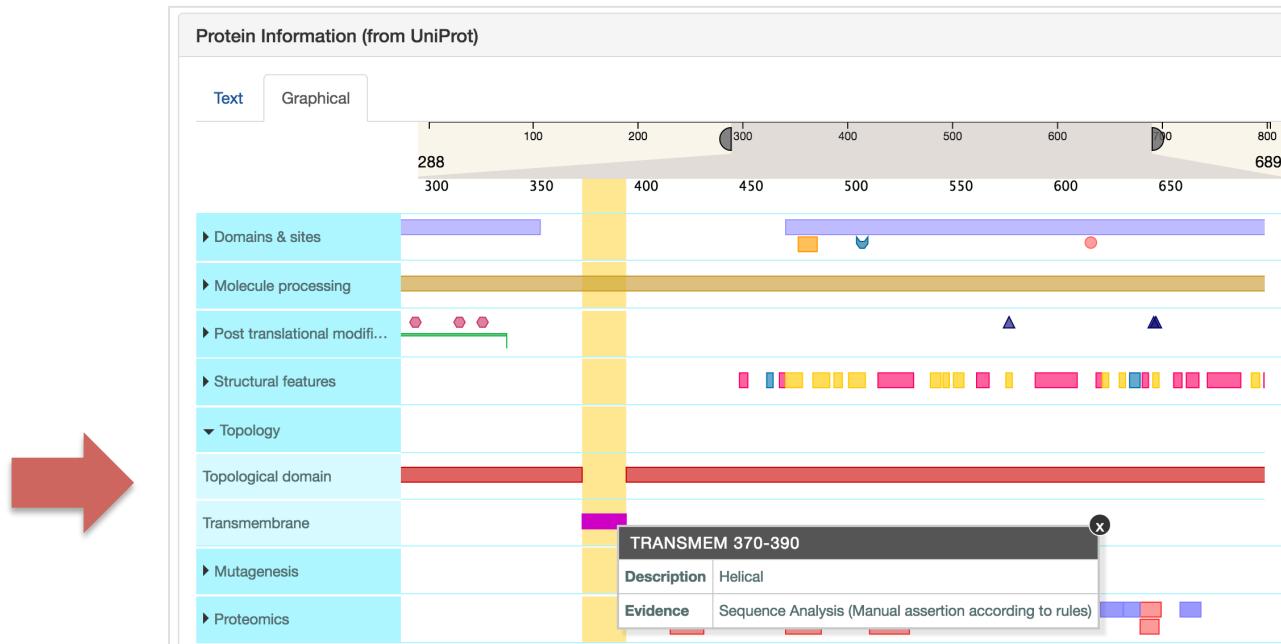
[Pathways](#)

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:



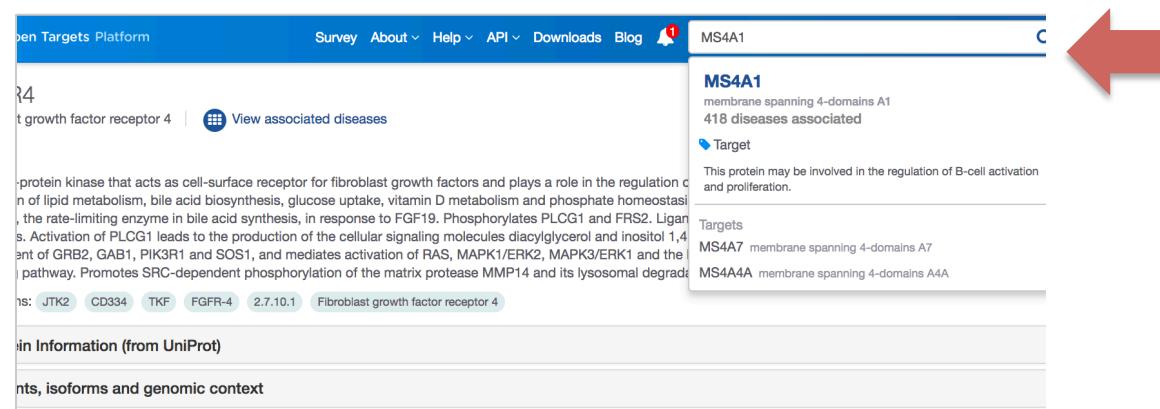
f) Click on the Protein information (from UniProt) tab. Now, click on the Graphical view option, then click on the Topology menu to see the annotated domains: extracellular, transmembrane and intracellular. The transmembrane (TM) domain goes from amino acid 370 to 390. Gene *FGFR4* codes for a receptor,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



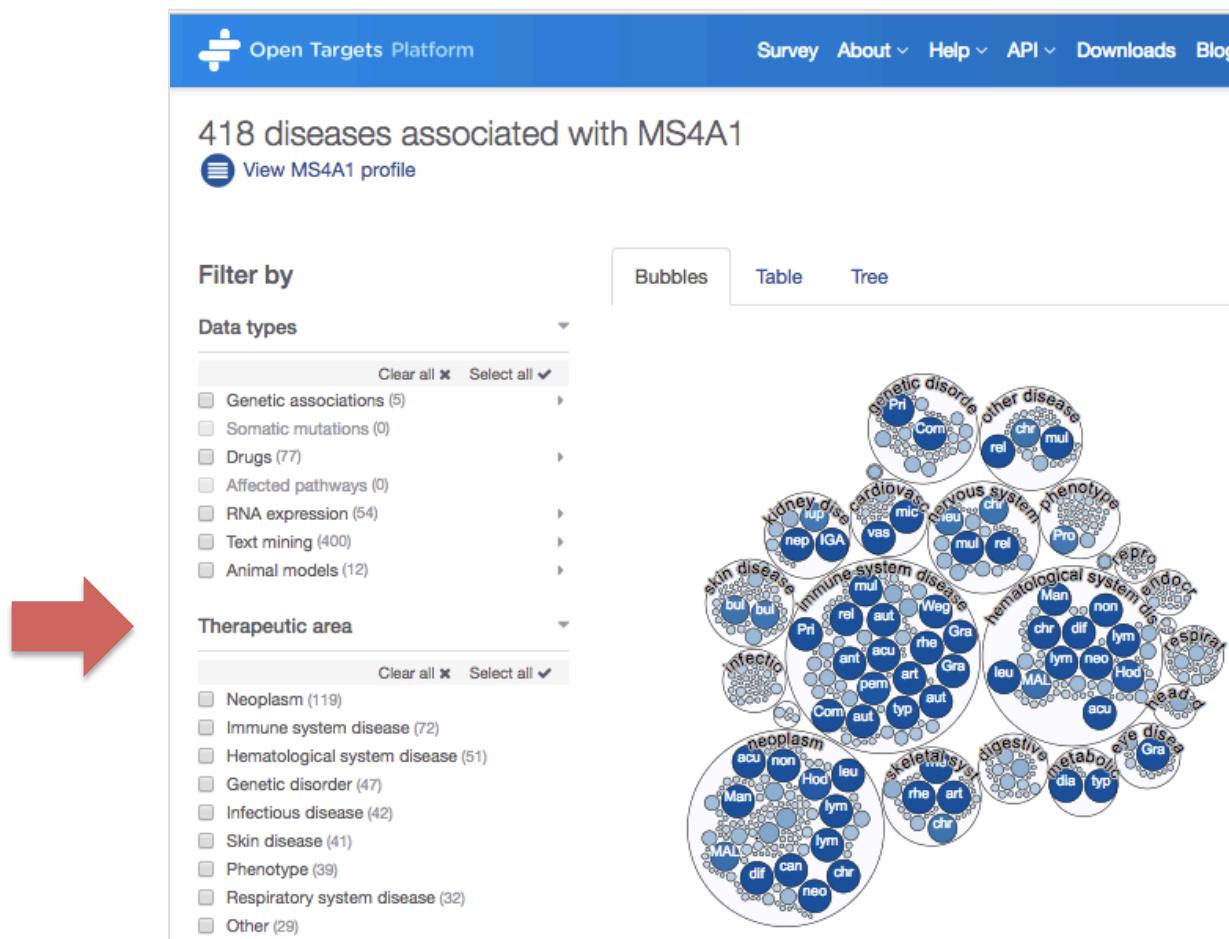
Exercise 2 – *MS4A1* as a possible drug target in the treatment of non-Hodgkin's lymphoma

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:

16 diseases associated with MS4A1

[View MS4A1 profile](#)

Filter by

Data types

- Genetic associations (0)
 Somatic mutations (0)
 Drugs (16)
 Affected pathways (0)
 RNA expression (8)
 Text mining (50)
 Animal models (2)

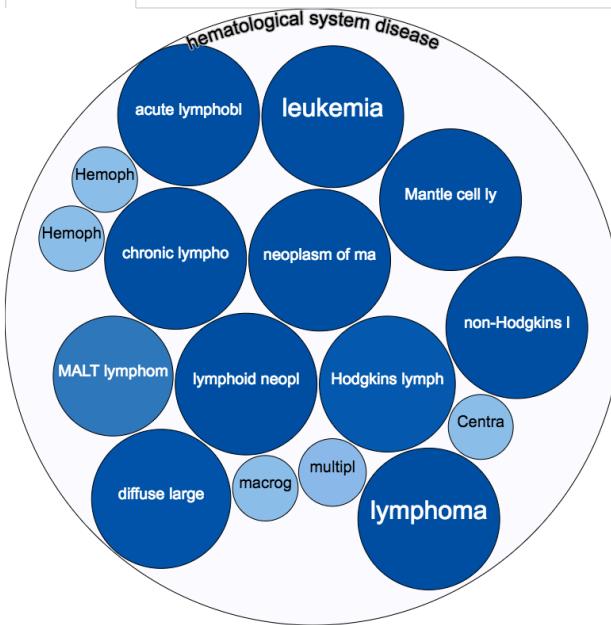
Therapeutic area

- Immune system disease (25)
 Neoplasm (19)
 Hematological system disease (16)
 Nervous system disease (9)
 Skin disease (6)
 Other (6)
 Digestive system disease (5)

Bubbles

Table

Tree



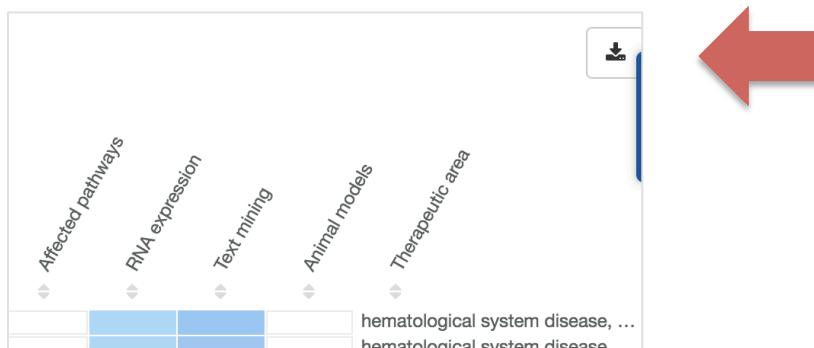
b) The default display for the diseases associated with a target is the Bubbles view.

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':

The figure shows a detailed view of drug information for Rituximab. The table has several columns: Disease (non-Hodgkins lymphoma), Drug (RITUXIMAB), Phase (Phase IV), Status (Completed), Type (Antibody), Mechanism of action (B-lymphocyte antigen CD20 inhibitor DailyMed), Activity (antagonist), Target class (CD20 Ca²⁺ channel family), and Evidence source (Curated from Clinical Trials Information). The 'Drugs' section header is visible above the table, along with links to 'Genetic associations' and 'Somatic mutations'. A search bar and a download icon are also present.

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.



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:

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...

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).

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

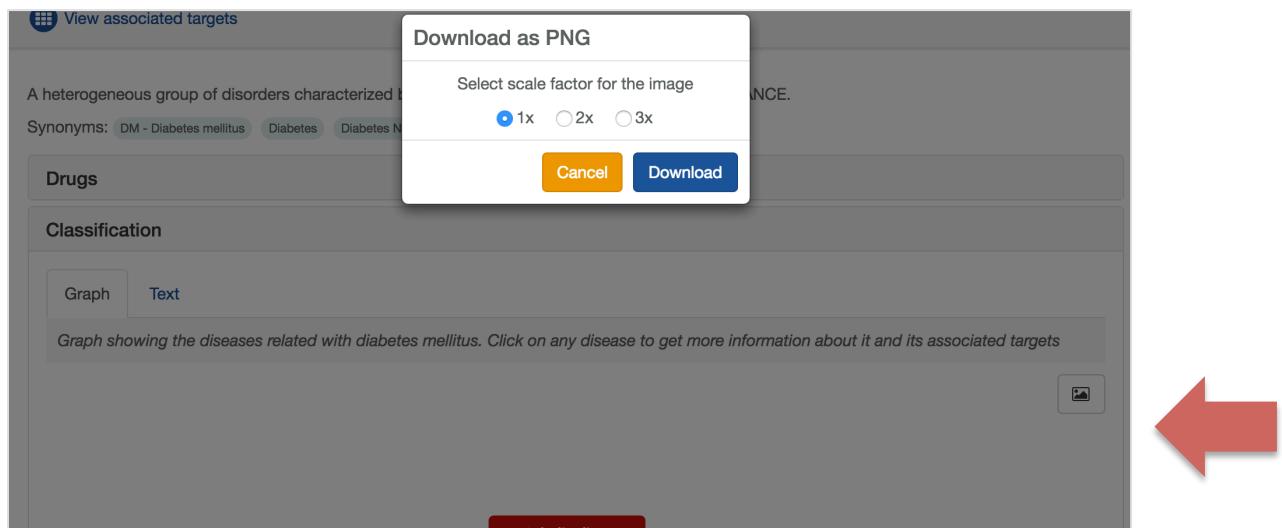
Drugs
Source: CHEMBL
Found 52 unique drugs: ABEMACICLIB ALEMTUZUMAB APITOLISIB AT-7519 BEXAROTENE BLINATUMOMAB BORTEZOMIB BRENTUXIMAB VEDOTIN BUPARLISIB CLOFARABINE COPANLISIB CYCLOSPORINE DASATINIB DEXAMETHASONE DEXAMETHASONE PHOSPHORIC ACID DOXORUBICIN DUVELISIB ENZASTAURIN EVEROLIMUS FILGRASTIM FLUDARABINE PHOSPHATE GALIXIMAB GSK-461364 IBRUTINIB IDELALISIB INTERFERON ALFA-2B IXAZOMIB CITRATE METHOTREXATE NILOTINIB NIVOLUMAB OBINUTUZUMAB OCRELIZUMAB OFATUMUMAB PALBOCICLIB PANOBINOSTAT PEGFILGRASTIM PICTILISIB PREDNISOLONE PREDNISONE RITUXIMAB ROMIDEPSIN ROMIPLOSTIM RP-6530 SIROLIMUS TASELISIB TEMSIROLIMUS THIOGUANINE TOSIUMOMAB VALSARTAN VENETOCLAX ...

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:

The screenshot shows the Open Targets Platform interface. At the top, there are navigation links for 'Downloads', 'Blog', and a notification bell icon with a red '1'. The search bar contains the query 'EGFR'. Below the search bar, the target name 'EGFR' is displayed in bold, followed by its definition 'epidermal growth factor receptor' and the statement '790 diseases associated'. A section titled 'Target' provides a detailed description of EGFR's function, mentioning it is a receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades. Known ligands include EGF, TGFA/TGF-alpha, amphiregulin, epigen/EPGN, BTC/betacellulin, epiregulin/EREG and HBEGF/heparin-binding EG...'. Below this, sections for 'Targets' and 'Diseases' list specific entries like 'CCDC50 coiled-coil domain containing 50' and 'ERBB2 erb-b2 receptor tyrosine kinase 2' under targets, and 'colorectal adenocarcinoma' and 'head and neck malignant neoplasia' under diseases. At the bottom of the main content area, there is a horizontal bar with various drug names: BLINATUMOMAB, DEXAMETHASONE, IBRUTINIB, OFATUMUMAB, SIROLIMUS, TASFIISIR, TFMSIROLIMUS, and THIOGUANINE.

You will then get to this page:

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

with all the disease associations with this target.

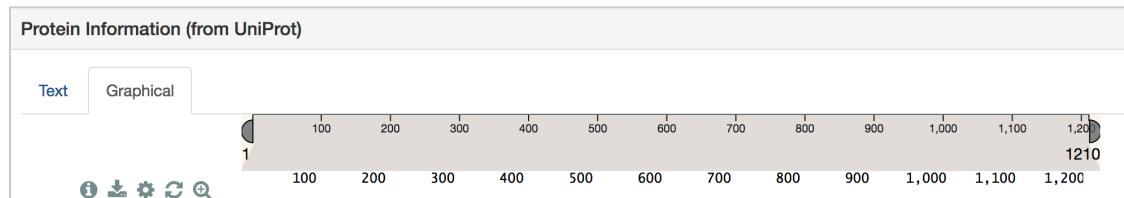
Click on 'View EGFR profile' link:

The screenshot shows the 'View EGFR profile' page. The header includes the Open Targets Platform logo and links for 'Survey' and 'About'. The main content area displays the text '790 diseases associated with EGFR' and a blue button labeled 'View EGFR profile' with a circular icon containing three horizontal lines.

You are now in the profile page of your target:

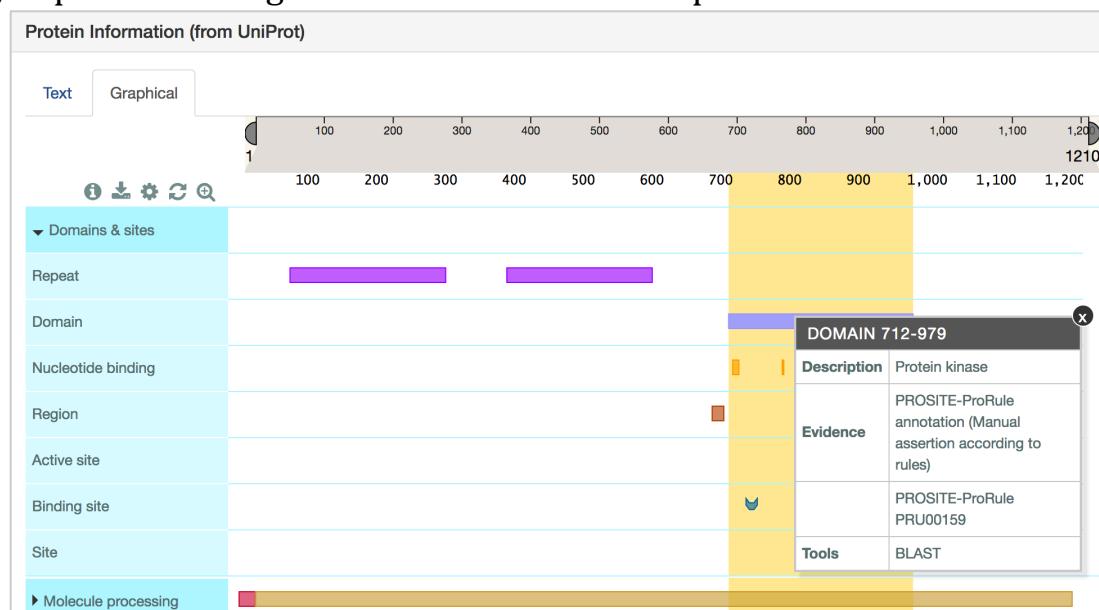
<http://targetvalidation.org/target/ENSG00000146648>

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”:



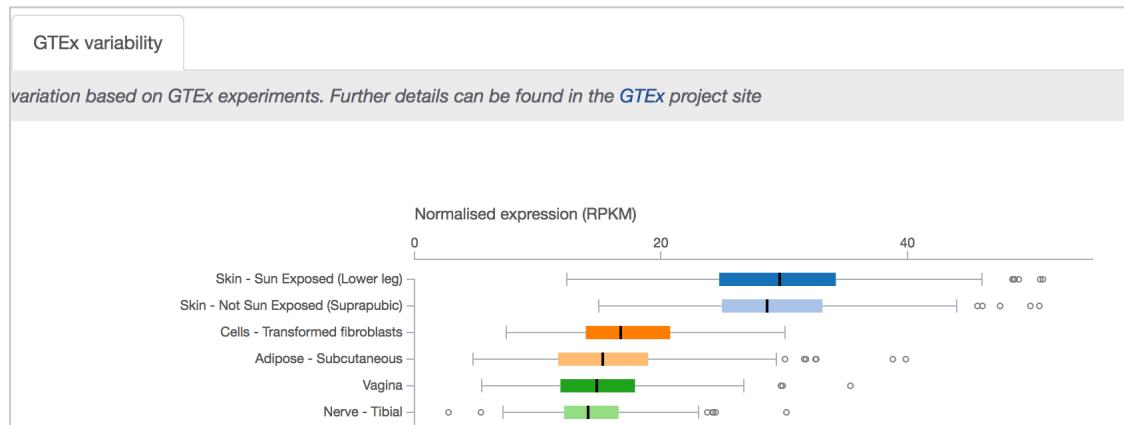
The kinase domain is at position 712-979 in the amino acid sequence. Other sites have been mapped to this protein, e.g. dimerization, phosphorylation and activation sites, ATP binding site and PIK3C2B interaction site.

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

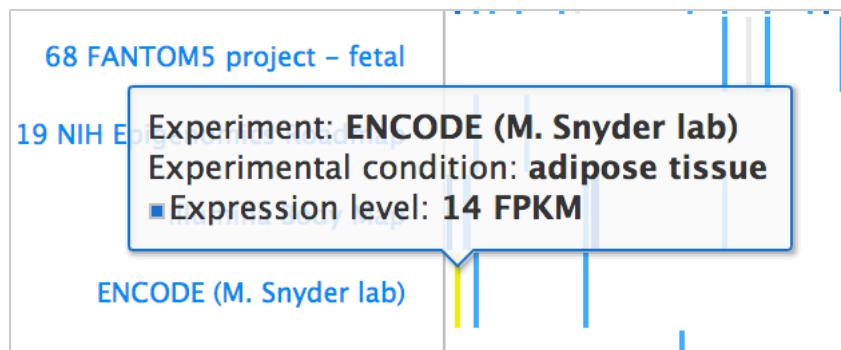


c) The tissues with the highest RNA baseline expression are:

1) skin from the GTEx project:



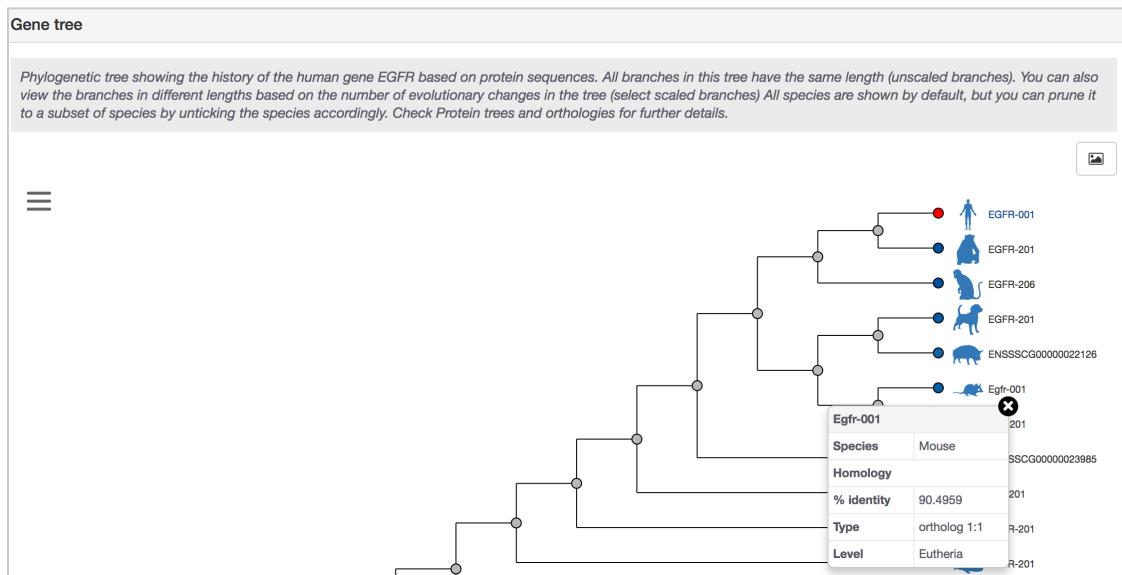
2) adipose tissue from the ENCODE project available in Expression Atlas:



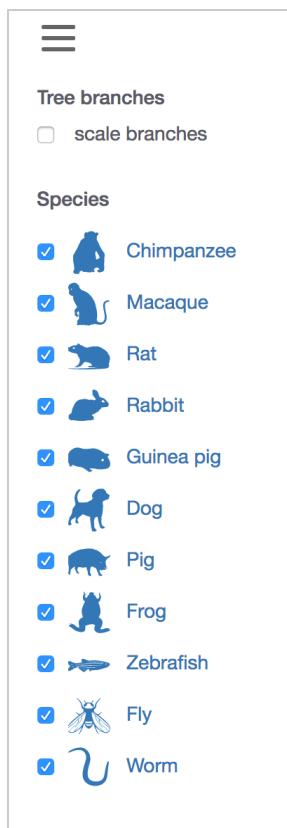
d) Expand the “Gene Ontology” tab to find out some of the molecular functions the *EGFR* protein e.g. actin filament binding, transmembrane receptor protein tyrosine kinase activity, MAP kinase kinase kinase activity, to name a few.

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 screenshot shows the ChEMBL Drugs search results for the query "IV". The results table has columns: Disease, Drug, Phase, Status, and Type. One row is shown for GEFITINIB, which targets non-small cell lung carcinoma, is in Phase IV, has a status of Terminated, and is a Small molecule.

Drug Information				
Disease	Drug	Phase	Status	Type
non-small cell lung carcinoma	GEFITINIB	Phase IV	Terminated	Small molecule

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