Precision medicine: NGS variant analysis and interpretation for translational research

Selecting the most relevant variants: How to filter

Fátima Al-Shahrour ● Javier Perales ● Elena Piñeiro

September 28, 2016

cnio Centro Nacional de Investigaciones Oncológicas

INSTITUTO GULBENKIAN DE CIÊNCIA

Additional annotations

chromosome	PolyPhen score	VEP's COSMIC ID
location	Condel effect	other IDs
mutation	Condel score	variation type
gene	SIFT effect	HGVS cDNA
feature	SIFT score	HGVS protein
feature type	gene HGNC	GMAF 1000 genomes
consequence	protein position	GMAF 1000 genomes percentage
	amino acids	ExAC percentage
PolyPhen effect	dbSNP ID	ExAC NFE percentage

Annotations from VEP

Additional annotations

chromosome	PolyPhen score	VEP's COSMIC ID	COSMIC original ID
location	Condel effect	other IDs	Pfam
mutation	Condel score	variation type	Uniprot
gene	SIFT effect	HGVS cDNA	Interpro
feature	SIFT score	HGVS protein	
feature type	gene HGNC	GMAF 1000 genomes	
consequence	protein position	GMAF 1000 genomes percentage	
APPRIS category	amino acids	ExAC percentage	
PolyPhen effect	dbSNP ID	ExAC NFE percentage	

Annotations from VEP

Enrichment of VEP annotations

Additional annotations

chromosome	PolyPhen score	VEP's COSMIC ID	COSMIC original ID	ClinVar disease
location	Condel effect	other IDs	Pfam	ClinVar clinical significance
mutation	Condel score	variation type	Uniprot	Homopolymer
gene	SIFT effect	HGVS cDNA	Interpro	Repeats
feature	SIFT score	HGVS protein	TumorPortal	CCLE gene
feature type	gene HGNC	GMAF 1000 genomes	Role of the gene in tumorgenesis	Frequency of gene in COSMIC
consequence	protein position	GMAF 1000 genomes percentage	KEGG data	Frequency of mutation in COSMIC
APPRIS category	amino acids	ExAC percentage	KEGG path ID	Consensual role
PolyPhen effect	dbSNP ID	ExAC NFE percentage	ClinVar ID	VSCORE

Annotations from VEP

Enrichment of VEP annotations

Annotations from other sources

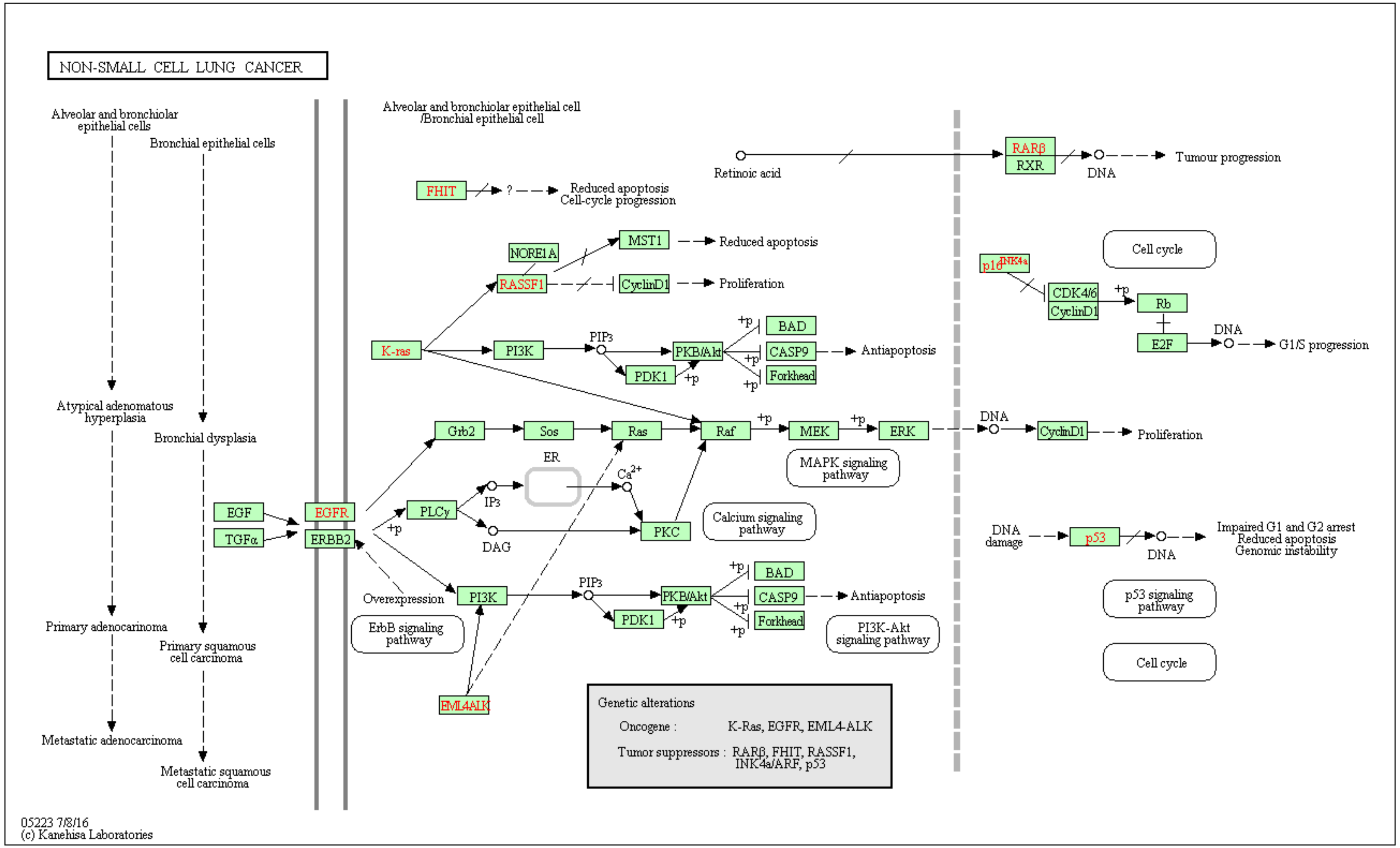
KEGG pathways

http://www.genome.jp/kegg/pathway.html

KEGG PATHWAY Database

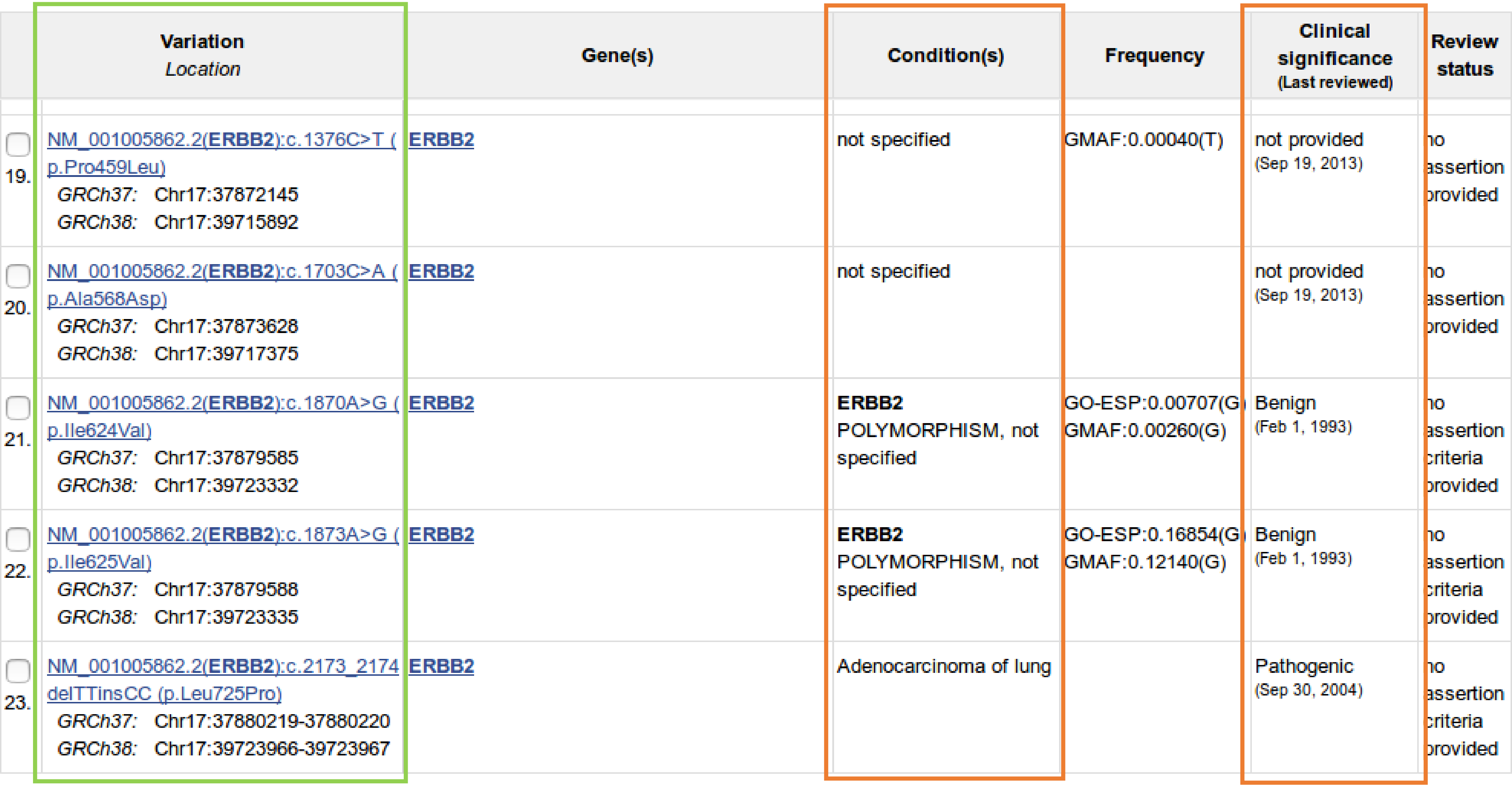
Wiring diagrams of molecular interactions, reactions, and relations

1. **Metabolism**
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human diseases**
7. **Drug development (structural relations between compounds)**


NON-SMALL CELL LUNG CANCER
Alveolar and bronchiolar epithelial cells
Bronchial epithelial cells
Atypical adenomatous hyperplasia
Bronchial dysplasia
Primary adenocarcinoma
Primary squamous cell carcinoma
Metastatic adenocarcinoma
Metastatic squamous cell carcinoma
Alveolar and bronchiolar epithelial cell /Bronchial epithelial cell
FHIT
?
Reduced apoptosis Cell-cycle progression
Retinoic acid
RARβ
RXR
DNA
Tumour progression
NORE1A
MST1
Reduced apoptosis
RASSF1
CyclinD1
Proliferation
K-ras
PI3K
PIP3
PKB/Akt
PDK1
BAD
CASP9
Forkhead
+p
Antiapoptosis
p16INK4a
CDK4/6
CyclinD1
Rb
E2F
DNA
G1/S progression
Cell cycle
Grb2
Sos
Ras
Raf
MEK
ERK
DNA
CyclinD1
Proliferation
MAPK signaling pathway
ER
Ca2+
IP3
DAG
PLCγ
PKC
Calcium signaling pathway
EGF
TGFα
EGFR
ERBB2
Overexpression
ErbB signaling pathway
PI3K
PIP3
PKB/Akt
PDK1
BAD
CASP9
Forkhead
Antiapoptosis
PI3K-Akt signaling pathway
EML4ALK
DNA damage
p53
DNA
Impaired G1 and G2 arrest Reduced apoptosis Genomic instability
p53 signaling pathway
Cell cycle
Genetic alterations
Oncogene : K-Ras, EGFR, EML4-ALK
Tumor suppressors : RARβ, FHIT, RASSF1, INK4a/ARF, p53
05223 7/8/16
(c) Kanehisa Laboratories

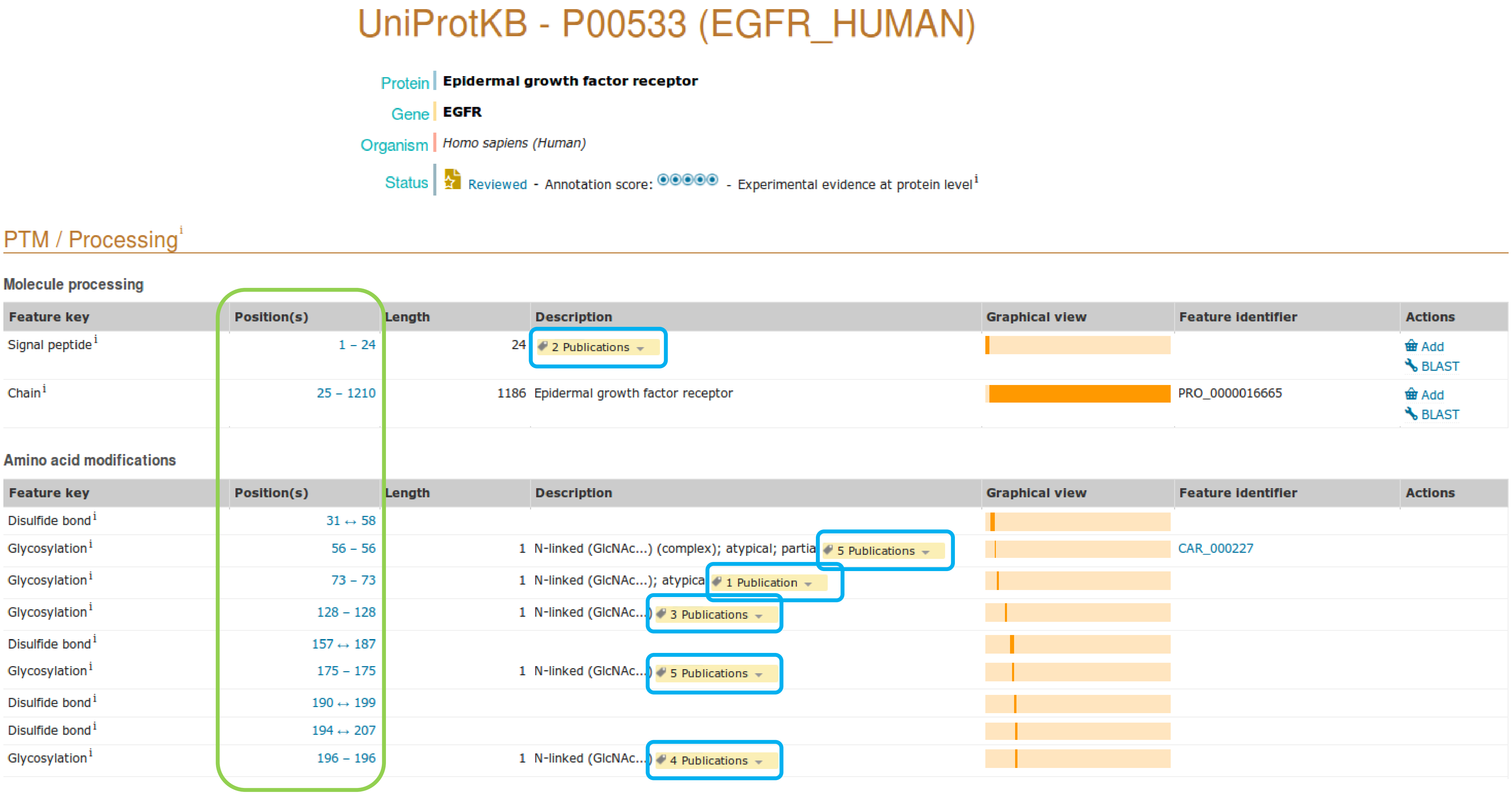
ClinVar

http://www.ncbi.nlm.nih.gov/clinvar/



	Variation *Location*	**Gene(s)**	**Condition(s)**	**Frequency**	**Clinical significance** **(Last reviewed)**	**Review status**
□ 19.	NM_001005862.2(**ERBB2**):c.1376C>T (p.Pro459Leu) *GRCh37:* Chr17:37872145 *GRCh38:* Chr17:39715892	**ERBB2**	not specified	GMAF:0.00040(T)	not provided (Sep 19, 2013)	no assertion provided
□ 20.	NM_001005862.2(**ERBB2**):c.1703C>A (p.Ala568Asp) *GRCh37:* Chr17:37873628 *GRCh38:* Chr17:39717375	**ERBB2**	not specified		not provided (Sep 19, 2013)	no assertion provided
□ 21.	NM_001005862.2(**ERBB2**):c.1870A>G (p.Ile624Val) *GRCh37:* Chr17:37879585 *GRCh38:* Chr17:39723332	**ERBB2**	**ERBB2** POLYMORPHISM, not specified	GO-ESP:0.00707(G GMAF:0.00260(G)	Benign (Feb 1, 1993)	no assertion criteria provided
□ 22.	NM_001005862.2(**ERBB2**):c.1873A>G (p.Ile625Val) *GRCh37:* Chr17:37879588 *GRCh38:* Chr17:39723335	**ERBB2**	**ERBB2** POLYMORPHISM, not specified	GO-ESP:0.16854(G GMAF:0.12140(G)	Benign (Feb 1, 1993)	no assertion criteria provided
□ 23.	NM_001005862.2(**ERBB2**):c.2173_2174 delTTinsCC (p.Leu725Pro) *GRCh37:* Chr17:37880219-37880220 *GRCh38:* Chr17:39723966-39723967	**ERBB2**	Adenocarcinoma of lung		Pathogenic (Sep 30, 2004)	no assertion criteria provided

UniProt additional information



UniProtKB - P00533 (EGFR_HUMAN)

Protein	**Epidermal growth factor receptor**
Gene	**EGFR**
Organism	*Homo sapiens (Human)*
Status	Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

PTM / Processing[i]

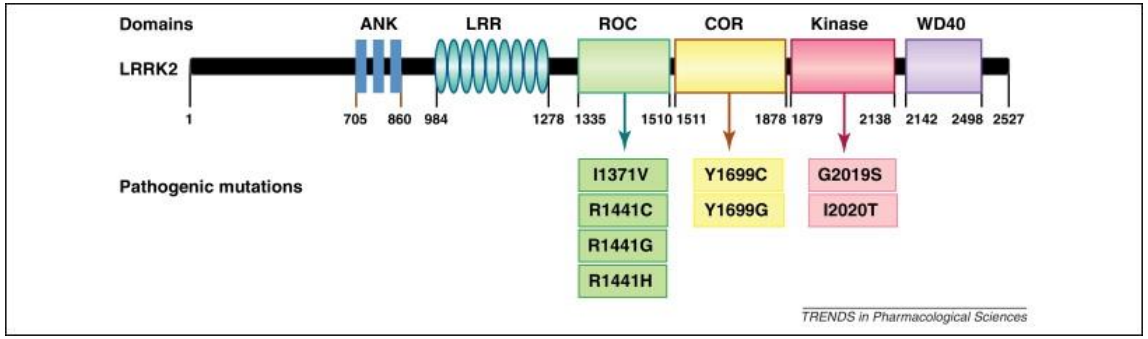
Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Signal peptide[i]	1 – 24	24	2 Publications			Add BLAST
Chain[i]	25 – 1210	1186	Epidermal growth factor receptor		PRO_0000016665	Add BLAST

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Disulfide bond[i]	31 ↔ 58					
Glycosylation[i]	56 – 56	1	N-linked (GlcNAc...) (complex); atypical; partia 5 Publications		CAR_000227	
Glycosylation[i]	73 – 73	1	N-linked (GlcNAc...); atypica 1 Publication			
Glycosylation[i]	128 – 128	1	N-linked (GlcNAc...) 3 Publications			
Disulfide bond[i]	157 ↔ 187					
Glycosylation[i]	175 – 175	1	N-linked (GlcNAc...) 5 Publications			
Disulfide bond[i]	190 ↔ 199					
Disulfide bond[i]	194 ↔ 207					
Glycosylation[i]	196 – 196	1	N-linked (GlcNAc...) 4 Publications			

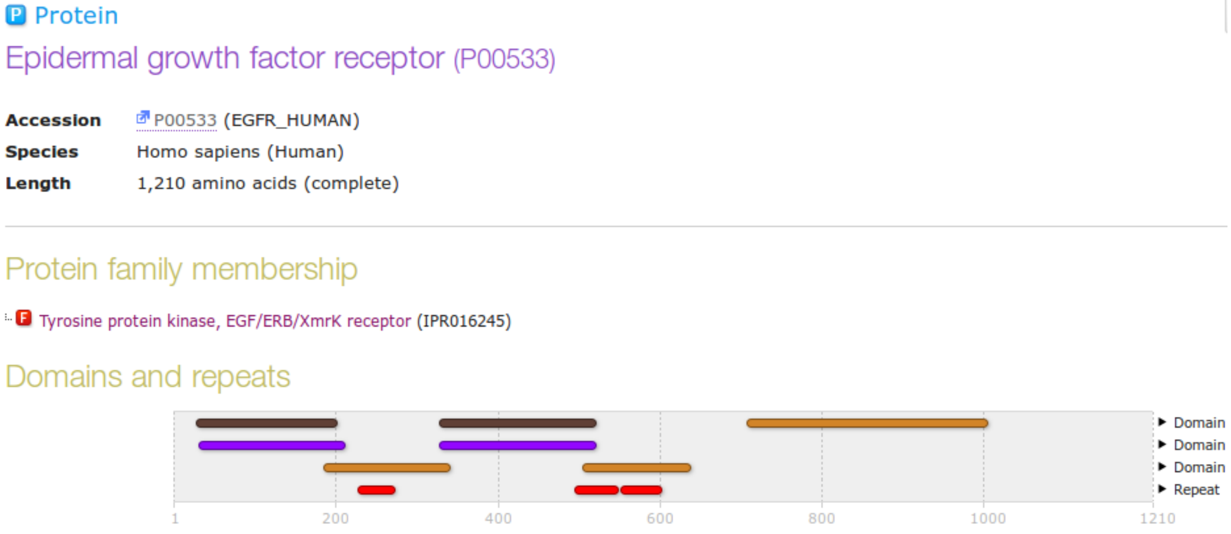
Additional domains information: pfam and Interpro



TRENDS in Pharmacological Sciences

EMBL-EBI InterPro Protein sequence analysis & classification

EMBL-EBI Pfam



Protein

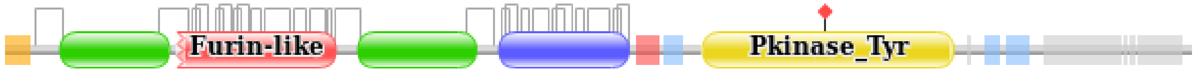
Epidermal growth factor receptor (P00533)

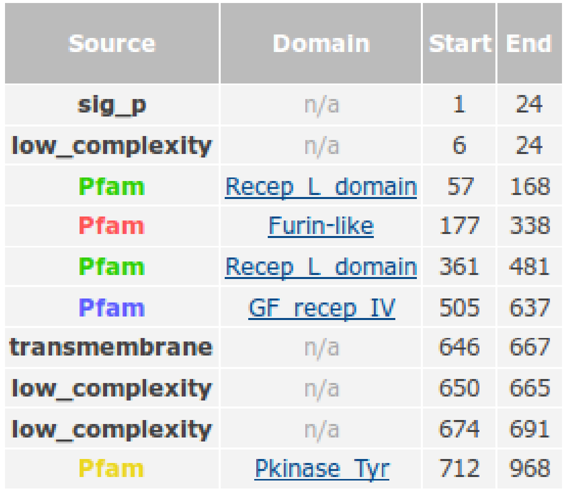
Accession	P00533 (EGFR_HUMAN)
Species	Homo sapiens (Human)
Length	1,210 amino acids (complete)

Protein family membership

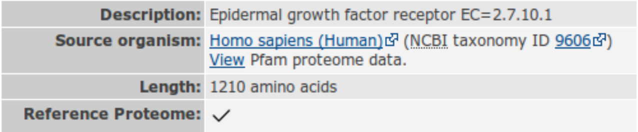
F Tyrosine protein kinase, EGF/ERB/XmrK receptor (IPR016245)

Domains and repeats





Source	Domain	Start	End
sig_p	n/a	1	24
low_complexity	n/a	6	24
Pfam	Recep_L_domain	57	168
Pfam	Furin-like	177	338
Pfam	Recep_L_domain	361	481
Pfam	GF_recep_IV	505	637
transmembrane	n/a	646	667
low_complexity	n/a	650	665
low_complexity	n/a	674	691
Pfam	Pkinase_Tyr	712	968



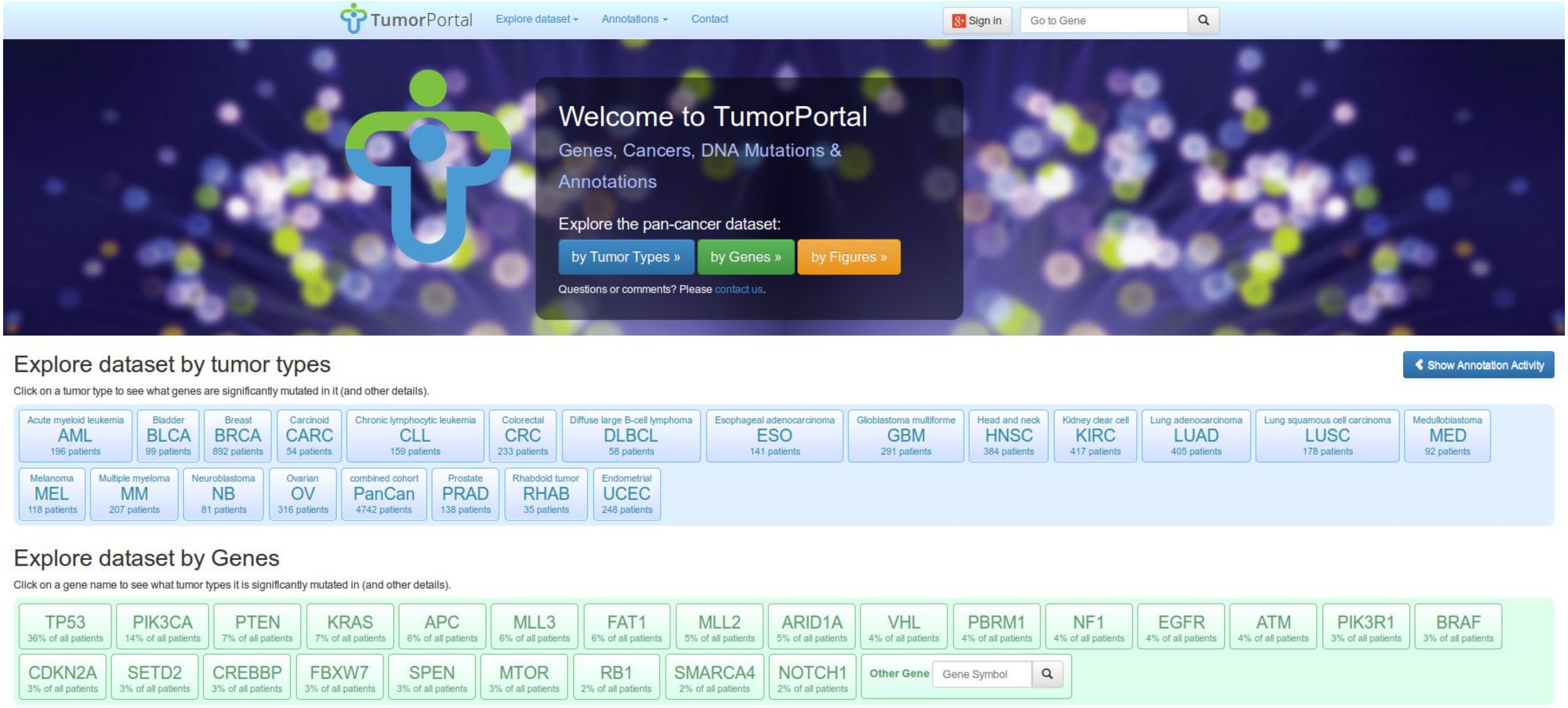
Description:	Epidermal growth factor receptor EC=2.7.10.1
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	1210 amino acids
Reference Proteome:	✓

Specific cancer information

- Relevance of the gene in carcinogenesis
  - TumorPortal
  - CCLE
  - COSMIC
- Frequency of the variant / gene in cancer
  - COSMIC
- Role of the gene in carcinogenesis (Oncogene or Tumor Suppressor)
  - COSMIC
  - oncodriveROLE

TumorPortal

http://www.tumorportal.org/



Welcome to TumorPortal

Genes, Cancers, DNA Mutations & Annotations

Explore the pan-cancer dataset:

by Tumor Types » by Genes » by Figures »

Questions or comments? Please contact us.

Explore dataset by tumor types

Show Annotation Activity

Click on a tumor type to see what genes are significantly mutated in it (and other details).

Acute myeloid leukemia AML 196 patients
Bladder BLCA 99 patients
Breast BRCA 892 patients
Carcinoid CARC 54 patients
Chronic lymphocytic leukemia CLL 159 patients
Colorectal CRC 233 patients
Diffuse large B-cell lymphoma DLBCL 58 patients
Esophageal adenocarcinoma ESO 141 patients
Glioblastoma multiforme GBM 291 patients
Head and neck HNSC 384 patients
Kidney clear cell KIRC 417 patients
Lung adenocarcinoma LUAD 405 patients
Lung squamous cell carcinoma LUSC 178 patients
Medulloblastoma MED 92 patients
Melanoma MEL 118 patients
Multiple myeloma MM 207 patients
Neuroblastoma NB 81 patients
Ovarian OV 316 patients
combined cohort PanCan 4742 patients
Prostate PRAD 138 patients
Rhabdoid tumor RHAB 35 patients
Endometrial UCEC 248 patients

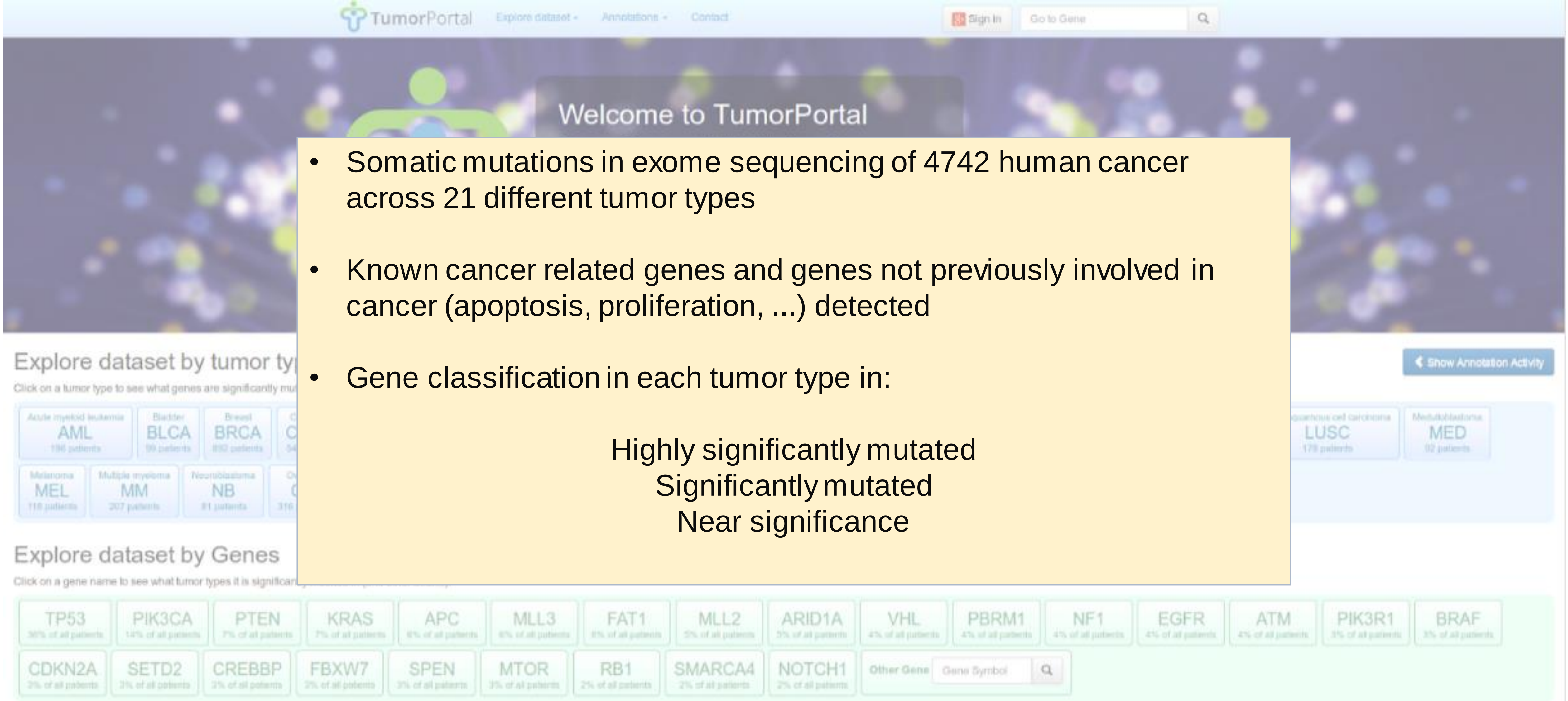
Explore dataset by Genes

Click on a gene name to see what tumor types it is significantly mutated in (and other details).

TP53 36% of all patients
PIK3CA 14% of all patients
PTEN 7% of all patients
KRAS 7% of all patients
APC 6% of all patients
MLL3 6% of all patients
FAT1 6% of all patients
MLL2 5% of all patients
ARID1A 5% of all patients
VHL 4% of all patients
PBRM1 4% of all patients
NF1 4% of all patients
EGFR 4% of all patients
ATM 4% of all patients
PIK3R1 3% of all patients
BRAF 3% of all patients
CDKN2A 3% of all patients
SETD2 3% of all patients
CREBBP 3% of all patients
FBXW7 3% of all patients
SPEN 3% of all patients
MTOR 3% of all patients
RB1 2% of all patients
SMARCA4 2% of all patients
NOTCH1 2% of all patients

Other Gene Gene Symbol

TumorPortal



Welcome to TumorPortal

Explore dataset by tumor ty

Click on a tumor type to see what genes are significantly mu

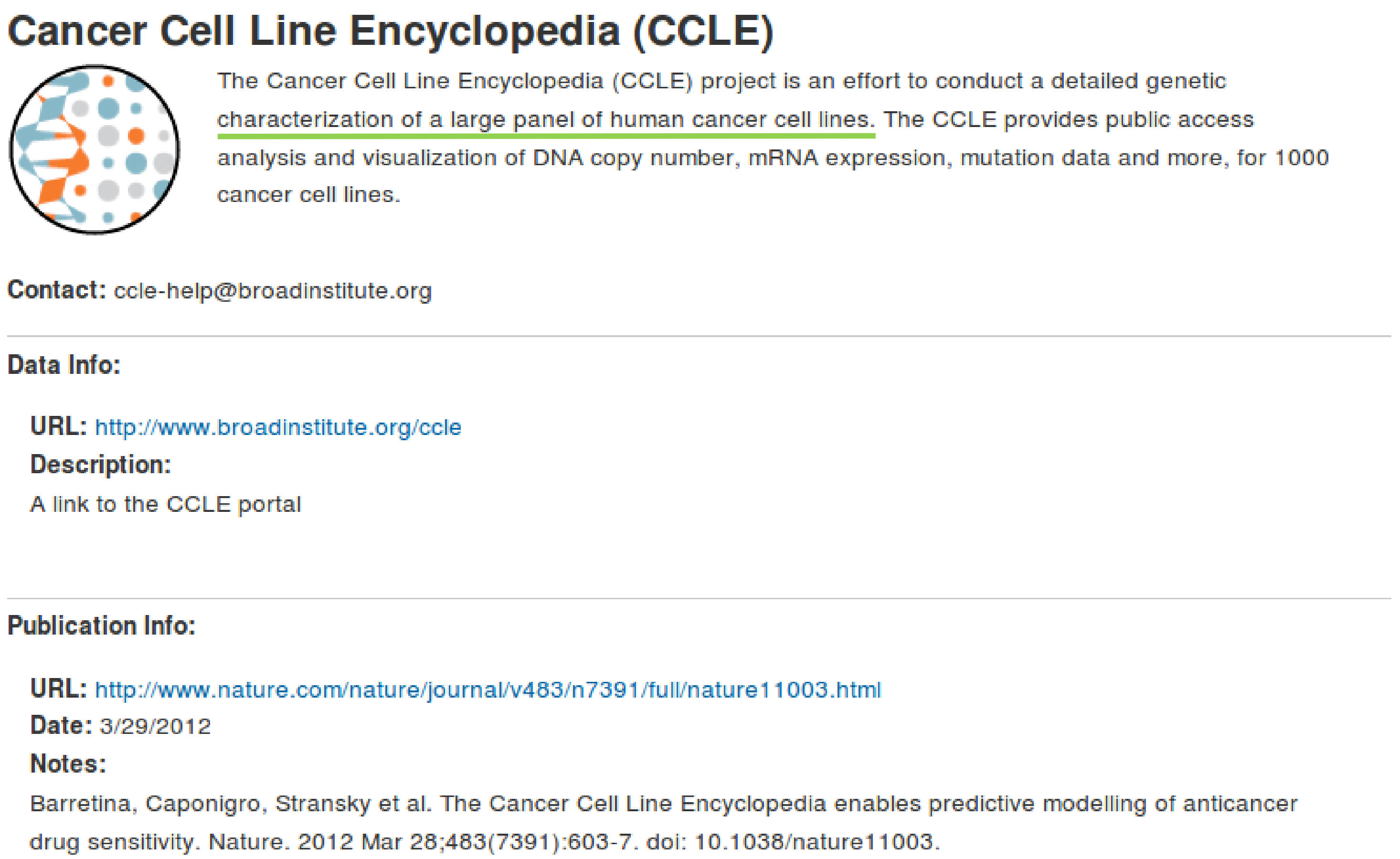
Explore dataset by Genes

Click on a gene name to see what tumor types it is significan

- Somatic mutations in exome sequencing of 4742 human cancer across 21 different tumor types
- Known cancer related genes and genes not previously involved in cancer (apoptosis, proliferation, ...) detected
- Gene classification in each tumor type in:

Highly significantly mutated
Significantly mutated
Near significance

CCLE



Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) project is an effort to conduct a detailed genetic characterization of a large panel of human cancer cell lines. The CCLE provides public access analysis and visualization of DNA copy number, mRNA expression, mutation data and more, for 1000 cancer cell lines.

Contact: ccle-help@broadinstitute.org

Data Info:

URL: http://www.broadinstitute.org/ccle
Description:
A link to the CCLE portal

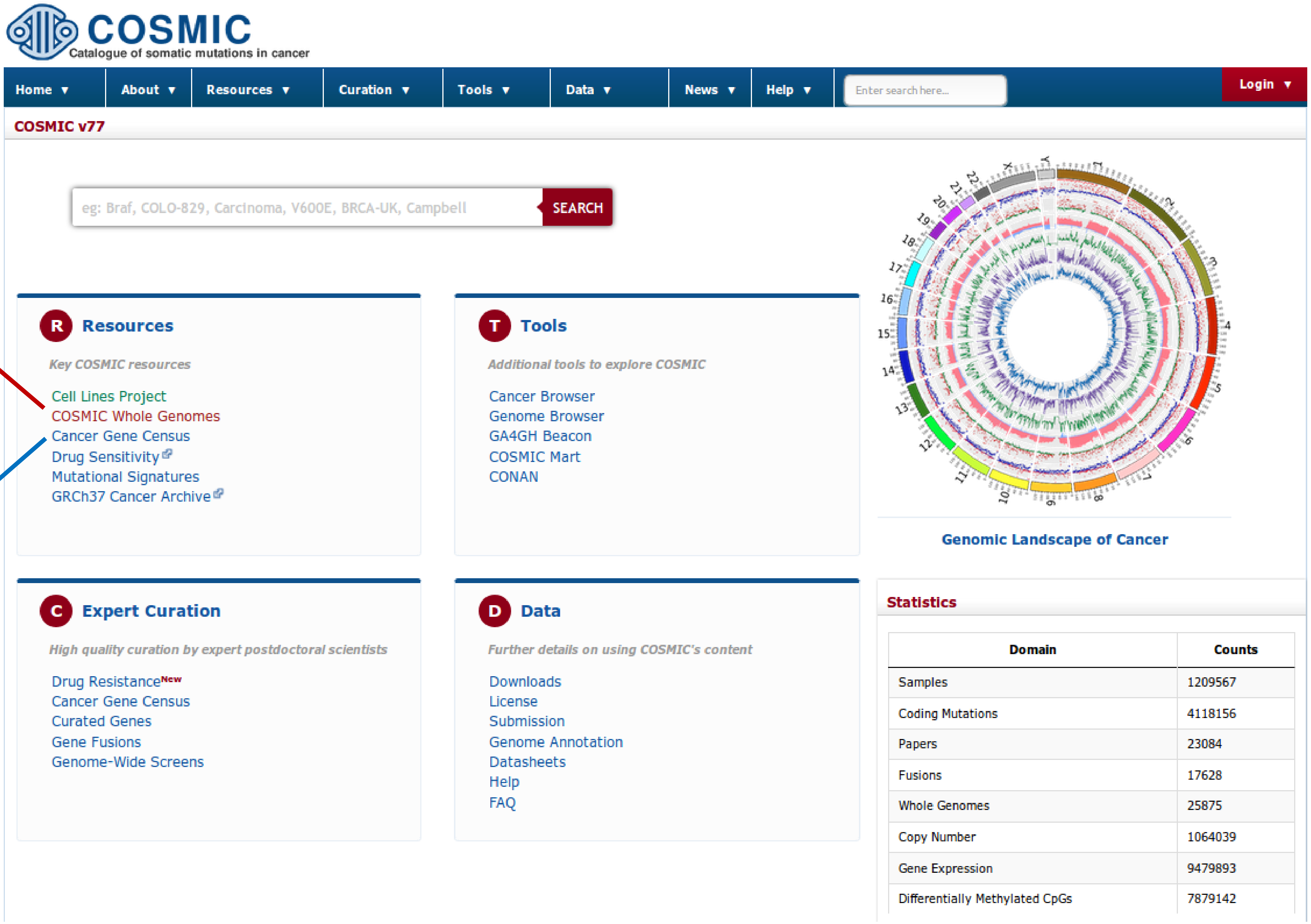
Publication Info:

URL: http://www.nature.com/nature/journal/v483/n7391/full/nature11003.html
Date: 3/29/2012
Notes:
Barretina, Caponigro, Stransky et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

COSMIC additional information

variant / gene frequency in database

cancer genes information (e.g. role of the gene)



COSMIC v77

R Resources

Key COSMIC resources

- Cell Lines Project
- COSMIC Whole Genomes
- Cancer Gene Census
- Drug Sensitivity
- Mutational Signatures
- GRCh37 Cancer Archive

T Tools

Additional tools to explore COSMIC

- Cancer Browser
- Genome Browser
- GA4GH Beacon
- COSMIC Mart
- CONAN

Genomic Landscape of Cancer

C Expert Curation

High quality curation by expert postdoctoral scientists

- Drug Resistance New
- Cancer Gene Census
- Curated Genes
- Gene Fusions
- Genome-Wide Screens

D Data

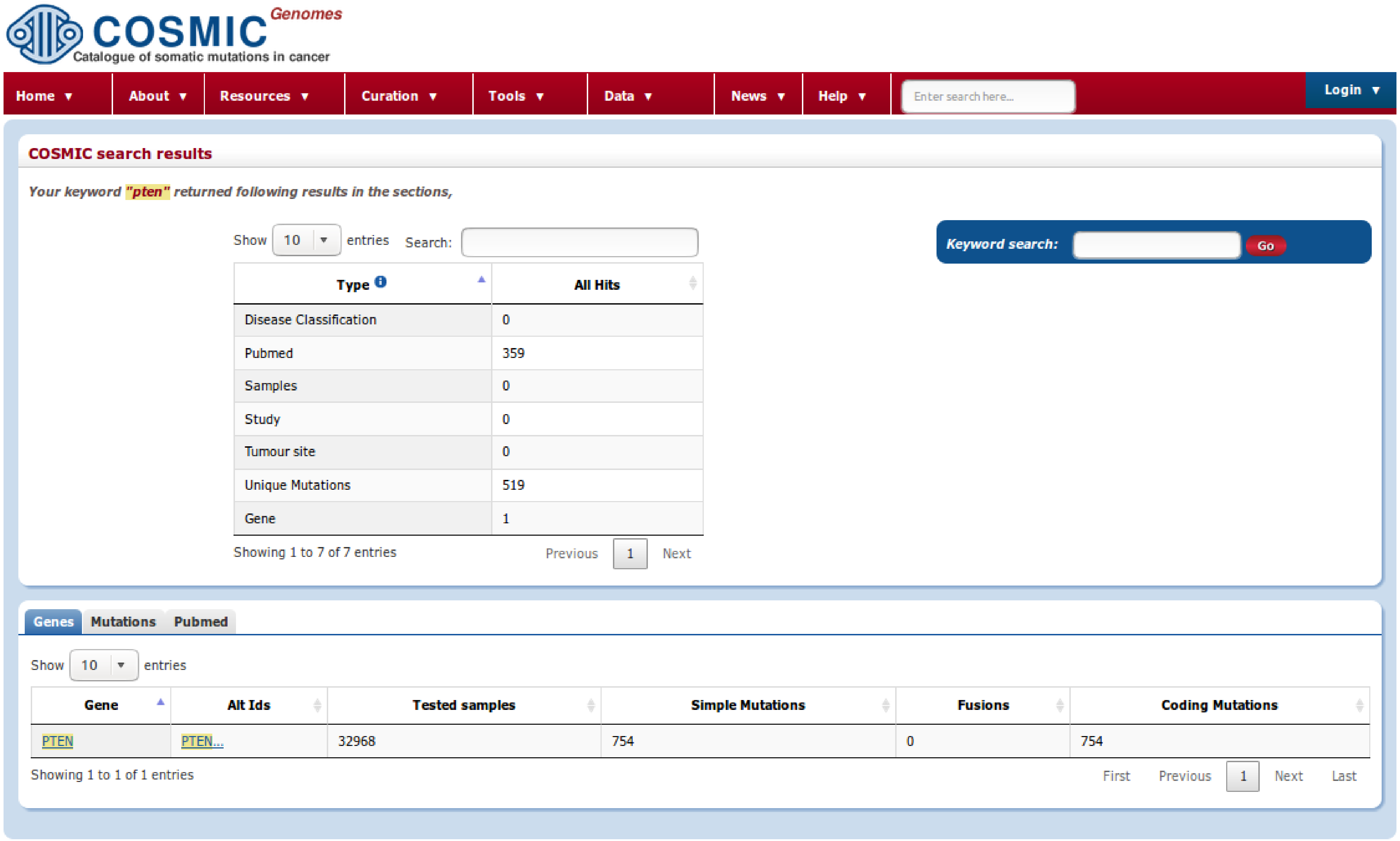
Further details on using COSMIC's content

- Downloads
- License
- Submission
- Genome Annotation
- Datasheets
- Help
- FAQ

Statistics

Domain	Counts
Samples	1209567
Coding Mutations	4118156
Papers	23084
Fusions	17628
Whole Genomes	25875
Copy Number	1064039
Gene Expression	9479893
Differentially Methylated CpGs	7879142

Variant / gene frequency in COSMIC



COSMIC search results

Your keyword "pten" returned following results in the sections,

Show 10 entries Search:

Type	All Hits
Disease Classification	0
Pubmed	359
Samples	0
Study	0
Tumour site	0
Unique Mutations	519
Gene	1

Showing 1 to 7 of 7 entries Previous 1 Next

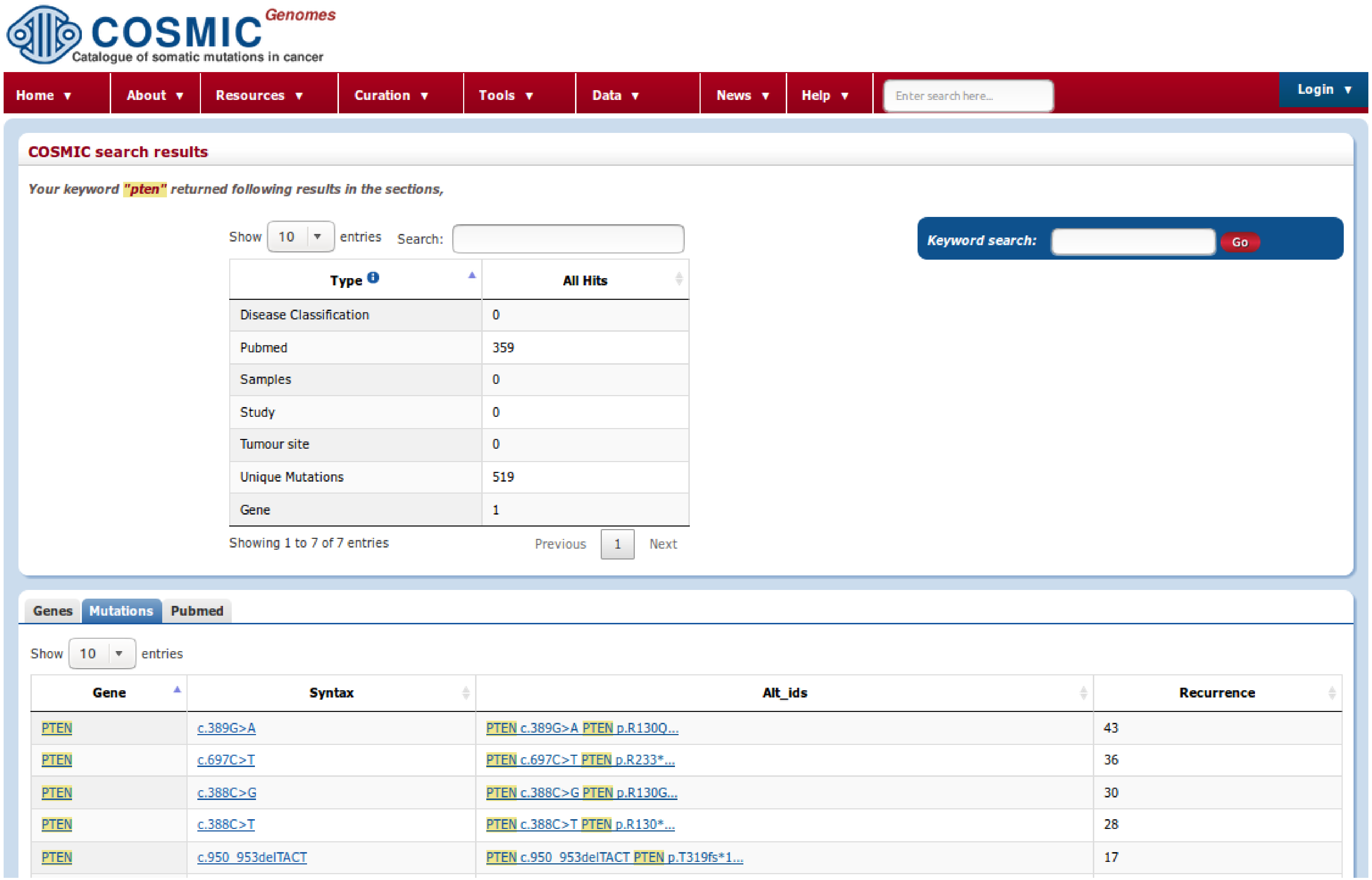
Genes Mutations Pubmed

Show 10 entries

Gene	Alt Ids	Tested samples	Simple Mutations	Fusions	Coding Mutations
PTEN	PTEN...	32968	754	0	754

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Variant / gene frequency in COSMIC



COSMIC search results

Your keyword "pten" returned following results in the sections,

Show 10 entries Search:

Type	All Hits
Disease Classification	0
Pubmed	359
Samples	0
Study	0
Tumour site	0
Unique Mutations	519
Gene	1

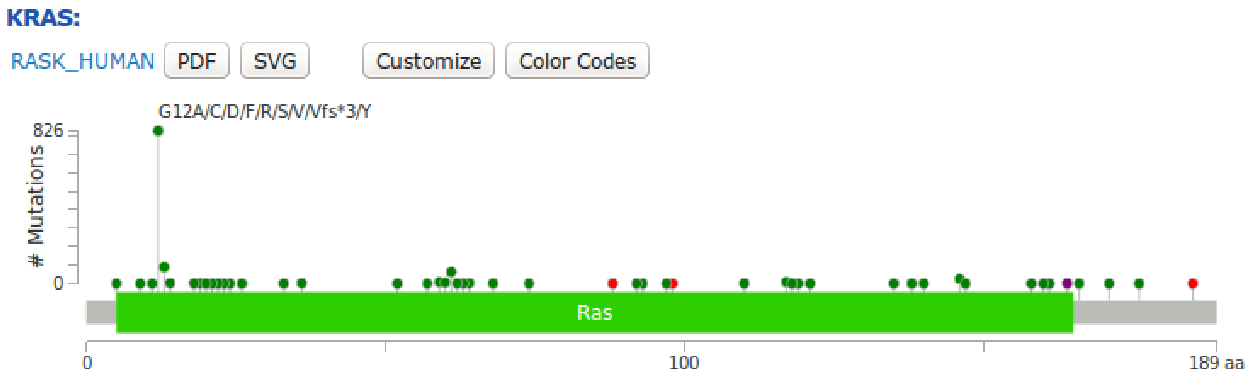
Showing 1 to 7 of 7 entries Previous 1 Next

Genes | Mutations | Pubmed

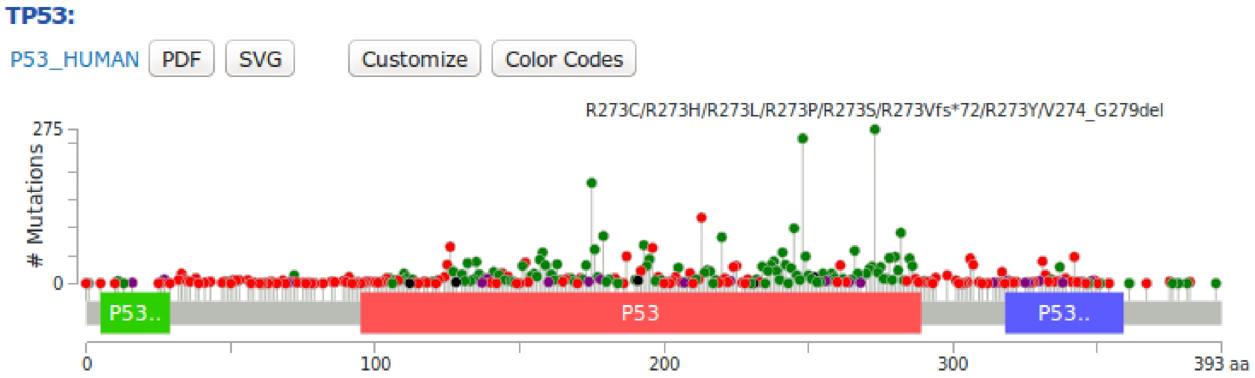
Show 10 entries

Gene	Syntax	Alt_ids	Recurrence
PTEN	c.389G>A	PTEN c.389G>A PTEN p.R130Q...	43
PTEN	c.697C>T	PTEN c.697C>T PTEN p.R233*...	36
PTEN	c.388C>G	PTEN c.388C>G PTEN p.R130G...	30
PTEN	c.388C>T	PTEN c.388C>T PTEN p.R130*...	28
PTEN	c.950_953delTACT	PTEN c.950_953delTACT PTEN p.T319fs*1...	17

Role of the gene as ONC or TSG



ONCOGENE



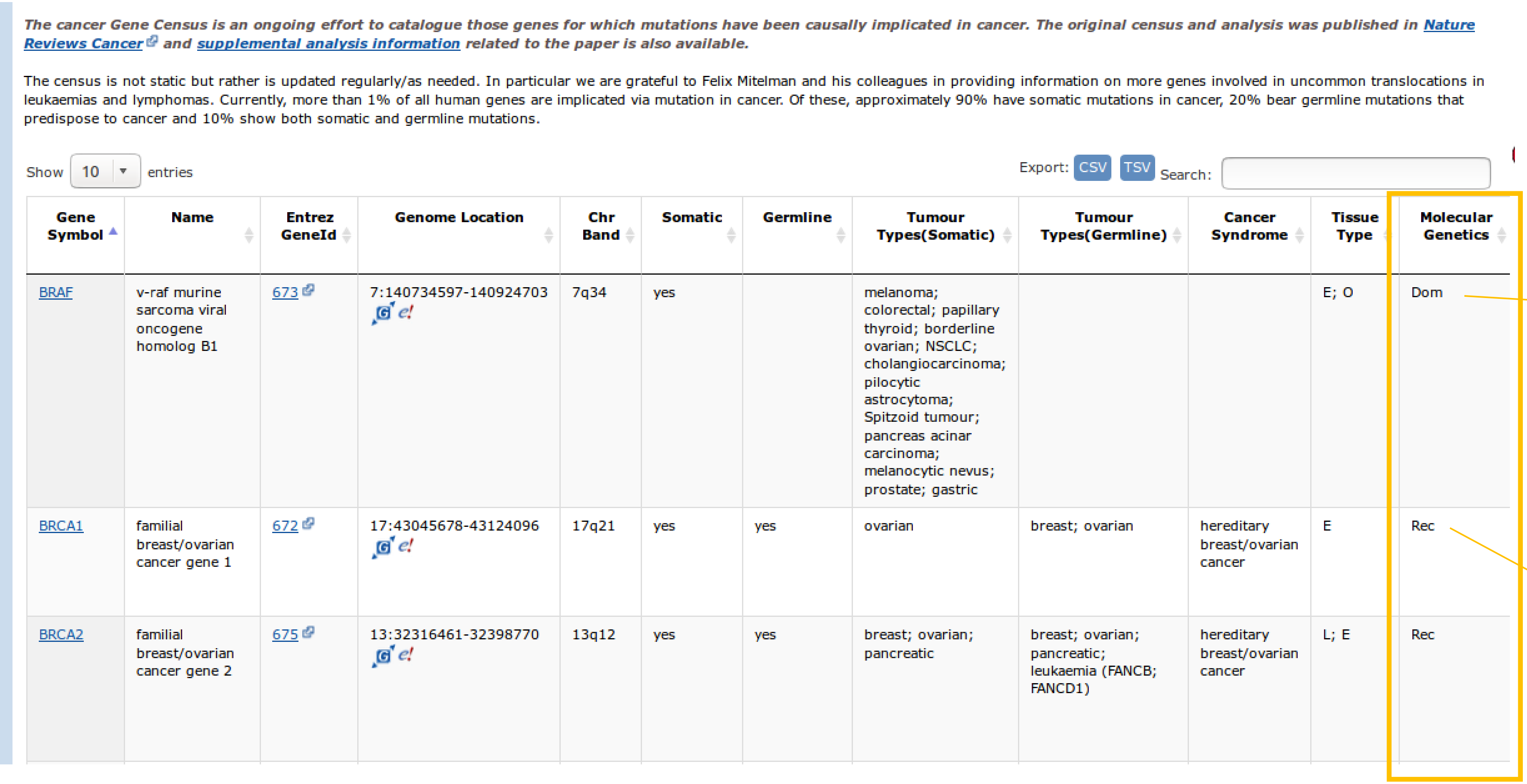
TUMOR SUPPRESSOR GENE

How frequencies can be interpreted?
How genes can be actioned for therapy?

Role of the gene as ONC or TSG

COSMIC
Catalogue of somatic mutations in cancer

http://cancer.sanger.ac.uk/census/



The cancer Gene Census is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. The original census and analysis was published in Nature Reviews Cancer and supplemental analysis information related to the paper is also available.

The census is not static but rather is updated regularly/as needed. In particular we are grateful to Felix Mitelman and his colleagues in providing information on more genes involved in uncommon translocations in leukaemias and lymphomas. Currently, more than 1% of all human genes are implicated via mutation in cancer. Of these, approximately 90% have somatic mutations in cancer, 20% bear germline mutations that predispose to cancer and 10% show both somatic and germline mutations.

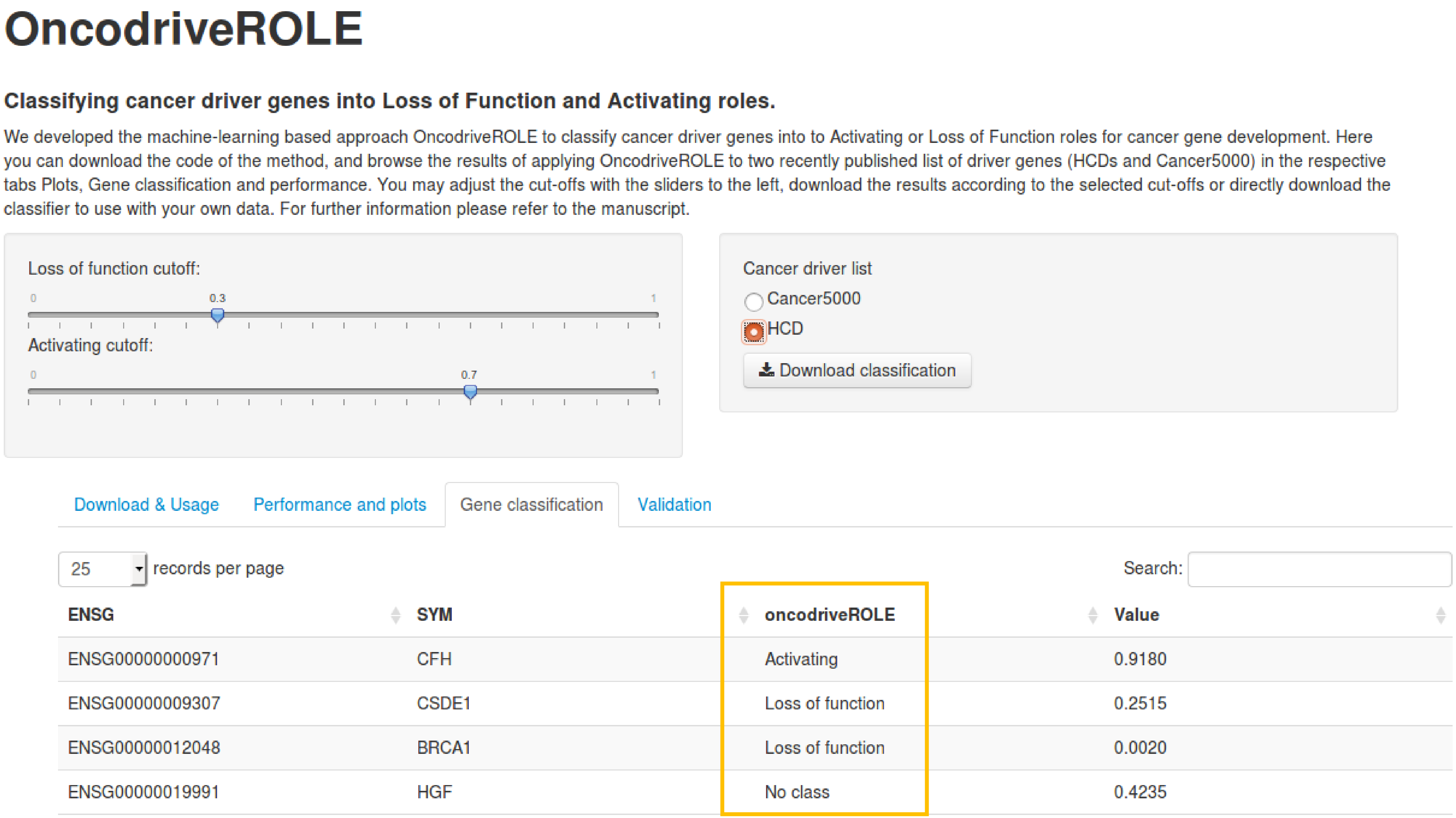
Gene Symbol	Name	Entrez GeneId	Genome Location	Chr Band	Somatic	Germline	Tumour Types(Somatic)	Tumour Types(Germline)	Cancer Syndrome	Tissue Type	Molecular Genetics
BRAF	v-raf murine sarcoma viral oncogene homolog B1	673	7:140734597-140924703	7q34	yes		melanoma; colorectal; papillary thyroid; borderline ovarian; NSCLC; cholangiocarcinoma; pilocytic astrocytoma; Spitzoid tumour; pancreas acinar carcinoma; melanocytic nevus; prostate; gastric			E; O	Dom
BRCA1	familial breast/ovarian cancer gene 1	672	17:43045678-43124096	17q21	yes	yes	ovarian	breast; ovarian	hereditary breast/ovarian cancer	E	Rec
BRCA2	familial breast/ovarian cancer gene 2	675	13:32316461-32398770	13q12	yes	yes	breast; ovarian; pancreatic	breast; ovarian; pancreatic; leukaemia (FANCB; FANCD1)	hereditary breast/ovarian cancer	L; E	Rec

ONC

TSG

Role of the gene as ONC or TSG

http://bg.upf.edu/oncodrive-role/



OncodriveROLE

Classifying cancer driver genes into Loss of Function and Activating roles.

We developed the machine-learning based approach OncodriveROLE to classify cancer driver genes into to Activating or Loss of Function roles for cancer gene development. Here you can download the code of the method, and browse the results of applying OncodriveROLE to two recently published list of driver genes (HCDs and Cancer5000) in the respective tabs Plots, Gene classification and performance. You may adjust the cut-offs with the sliders to the left, download the results according to the selected cut-offs or directly download the classifier to use with your own data. For further information please refer to the manuscript.

Loss of function cutoff: 0.3

Activating cutoff: 0.7

Cancer driver list

○ Cancer5000

● HCD

ENSG	SYM	oncodriveROLE	Value
ENSG00000000971	CFH	Activating	0.9180
ENSG00000009307	CSDE1	Loss of function	0.2515
ENSG00000012048	BRCA1	Loss of function	0.0020
ENSG00000019991	HGF	No class	0.4235

Filtering process

Once variants have been annotated we can remove the non likely relevant using the annotation information.

We can select manually those that seem more relevant according to a set of criteria.

A useful tool for the selection is the prioritization based on a score calculation computed from selected **annotations**. This provides a ranked list of variants with the most relevant at the top.

Components in the selection criteria and **score** calculation vary with the pathology or condition under study.

Filtering criteria

Remove **artifacts**

Possible artifacts

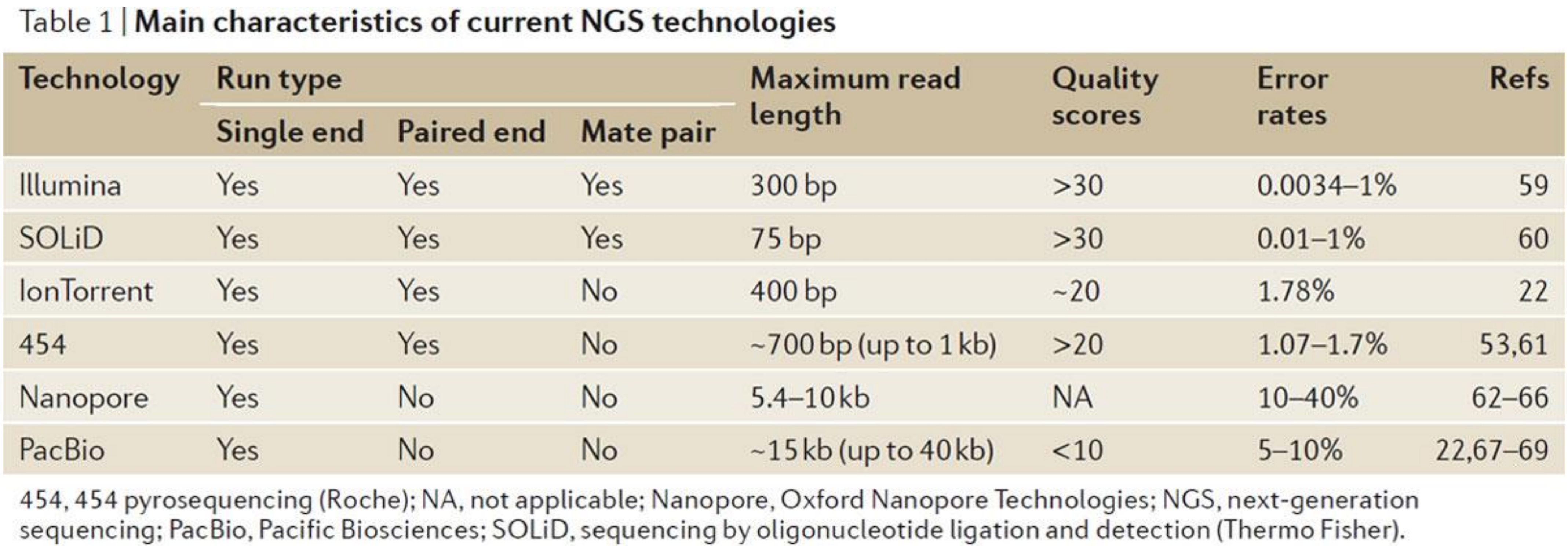


Table 1 | **Main characteristics of current NGS technologies**

Technology	Run type			Maximum read length	Quality scores	Error rates	Refs
	Single end	Paired end	Mate pair				
Illumina	Yes	Yes	Yes	300 bp	>30	0.0034–1%	59
SOLiD	Yes	Yes	Yes	75 bp	>30	0.01–1%	60
IonTorrent	Yes	Yes	No	400 bp	~20	1.78%	22
454	Yes	Yes	No	~700 bp (up to 1 kb)	>20	1.07–1.7%	53,61
Nanopore	Yes	No	No	5.4–10 kb	NA	10–40%	62–66
PacBio	Yes	No	No	~15 kb (up to 40 kb)	<10	5–10%	22,67–69

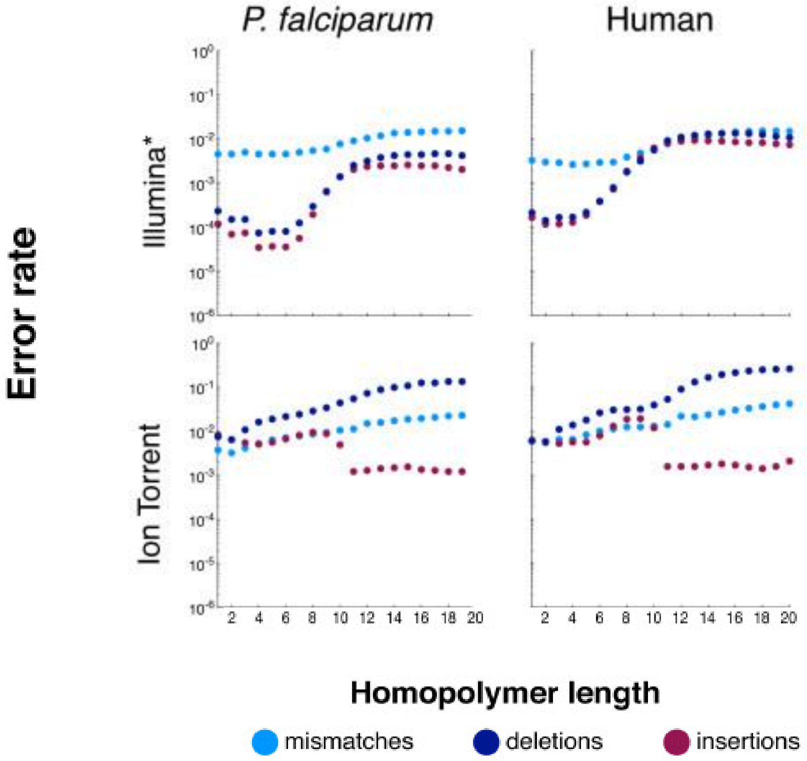
454, 454 pyrosequencing (Roche); NA, not applicable; Nanopore, Oxford Nanopore Technologies; NGS, next-generation sequencing; PacBio, Pacific Biosciences; SOLiD, sequencing by oligonucleotide ligation and detection (Thermo Fisher).

Nature Reviews Genetics 17,459–469(2016)doi:10.1038/nrg.2016.57

Sequencing strategies differ in different aspects as the error rates they produce and the kind of sequencing errors they introduce

Possible artifacts

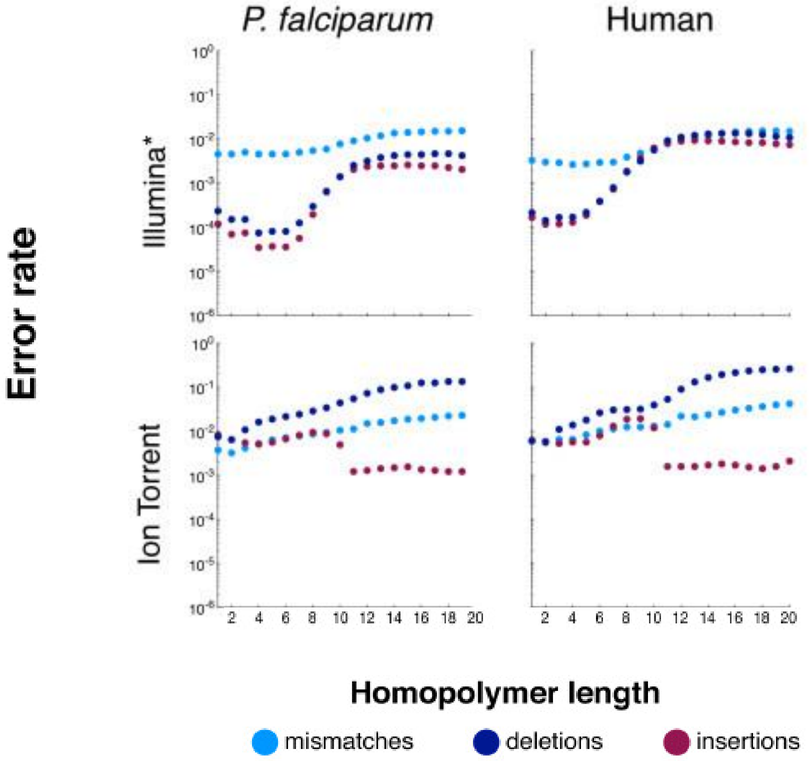
- Base-calling errors
  - Indel errors: Rare in Illumina. Main source of errors in IonTorrent and 454.
  - Substitution errors: Dominant in Illumina and SOLiD platforms.



Adapted from: Ross *et al. Genome Biology* 2013 **14**:R51 doi:10.1186/gb-2013-14-5-r51

Possible artifacts

- Base-calling errors
  - Indel errors: Rare in Illumina. Main source of errors in IonTorrent and 454.
  - Substitution errors: Dominant in Illumina and SOLiD platforms.



Check in IGV

Detection

Quality filters

Homopolymeric regions

Repetition in same technology output

- Repetition can indicate a polymorphism if it is present in at least a 1% of the population.
- Repetition can indicate a frequent cancer alteration if its presence is validated in multiple cancer samples.
- Otherwise, it can be an artifact (especially in genes acting as tumor suppressor)

Adapted from: Ross *et al. Genome Biology* 2013 **14**:R51 doi:10.1186/gb-2013-14-5-r51

Filtering criteria

Remove **artifacts**

- High number of repetitions without high frequency in population or in cancer samples.
- Indels located in homopolymeric regions in data from sensitive platforms to this artifact.
- Variants with very low coverage.

Remove variants located in **no functional genes**: BACs, pseudogenes, …

Remove **polymorfisms**: Population frequency in 1000 Genomes project, ExAC, … >= 1% (if not interested in germline information)

Filtering criteria

Keep variants with **relevant consequences at transcriptional level**:

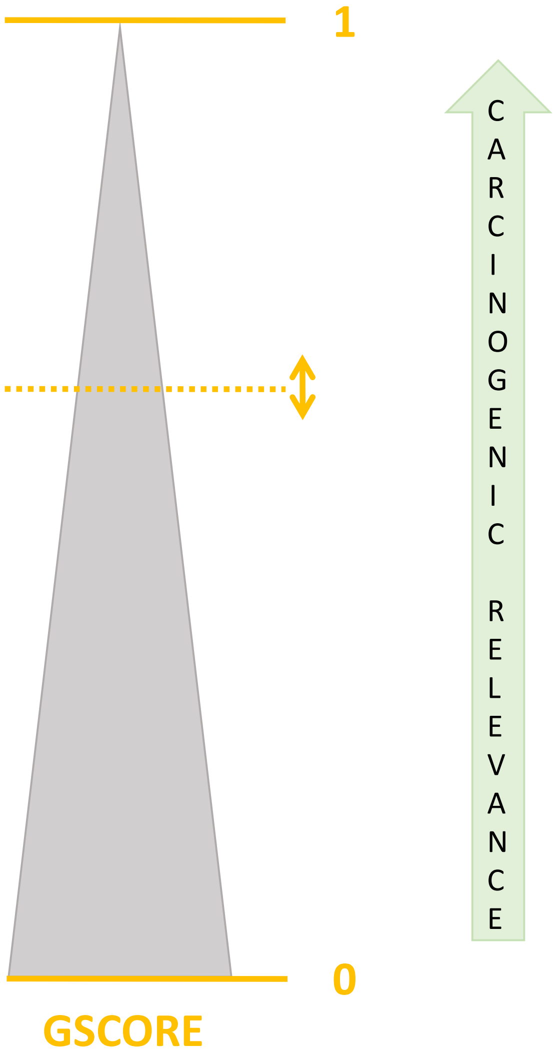
transcript_ablation | splice_donor_variant | splice_acceptor_variant | stop_gained | frameshift_variant | stop_lost | start_lost | transcript_amplification | inframe_insertion | inframe_deletion | missense_variant | protein_altering_variant | splice_region_variant | incomplete_terminal_codon_variant | stop_retained_variant

Keep variants with predicted **relevant consequence at protein level**: damaging in predictors, affecting domains

Keep variants with **clinical significance**: pathogenic ClinVar

Keep variants **relevant in the pathology**: pathogenic COSMIC, gene or variant frequently mutated in cancer, ...

Score calculation: an example



Feature	Value	Weight ONC	Weight TSG
Score prediction by PolyPhen	> 0.435	0.125/3	
Score prediction by Sift	<= 0.05	0.125/3	
Score prediction by CONDEL	> 0.468	0.125/3	
COSMIC	Pathogenic by FATHMM prediction	0.125/3	0.03125
Frequency of mutation in COSMIC	>= 100	0.125/3	
	< 100	(0.125 / 3) * (log(mutation frequency) / log(maximum mutation frequency))	
Frequency of gene in COSMIC	>= 100	0.125/3	0.03125
	< 100	(0.125 / 3) * (log(gene frequency) / log(maximum gene frequency))	0.03125 * (log(gene frequency) / log(maximum gene frequency))
VEP consequence	stop gain frameshift missense inframe insertion inframe deletion	0.125	
GMAF	< 1	0.125/2	
EXAC	< 1	0.125/2	
DOMAINS	Listed as relevant in cancer or previous last protein domain	0.125	
	Within a domain in other circumstances	0.125/2	
CLINVAR	Pathogenic	0.125	
ZYGOSITY	Homozygous	0.125	0.1875
ESSENTIALITY SCORE		0.125 * ES	

Impact

Pathogenicity

Frequencies

Impact

Frequencies

Impact

Pathogenicity

Impact