

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

November 15, 2017

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	gnomAD percentage
PolyPhen effect	dbSNP ID	gnomAD 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	gnomAD percentage	
PolyPhen effect	dbSNP ID	gnomAD 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 tumorigenesis	Frequency of gene in COSMIC
consequence	protein position	GMAF 1000 genomes percentage	KEGG data	Frequency of mutation in COSMIC
APPRIS category	amino acids	gnomAD percentage	KEGG path ID	Consensual role
PolyPhen effect	dbSNP ID	gnomAD NFE percentage	ClinVar ID	VSCORE

Annotations from VEP

Enrichment of VEP annotations

Annotations from other sources

Pathways

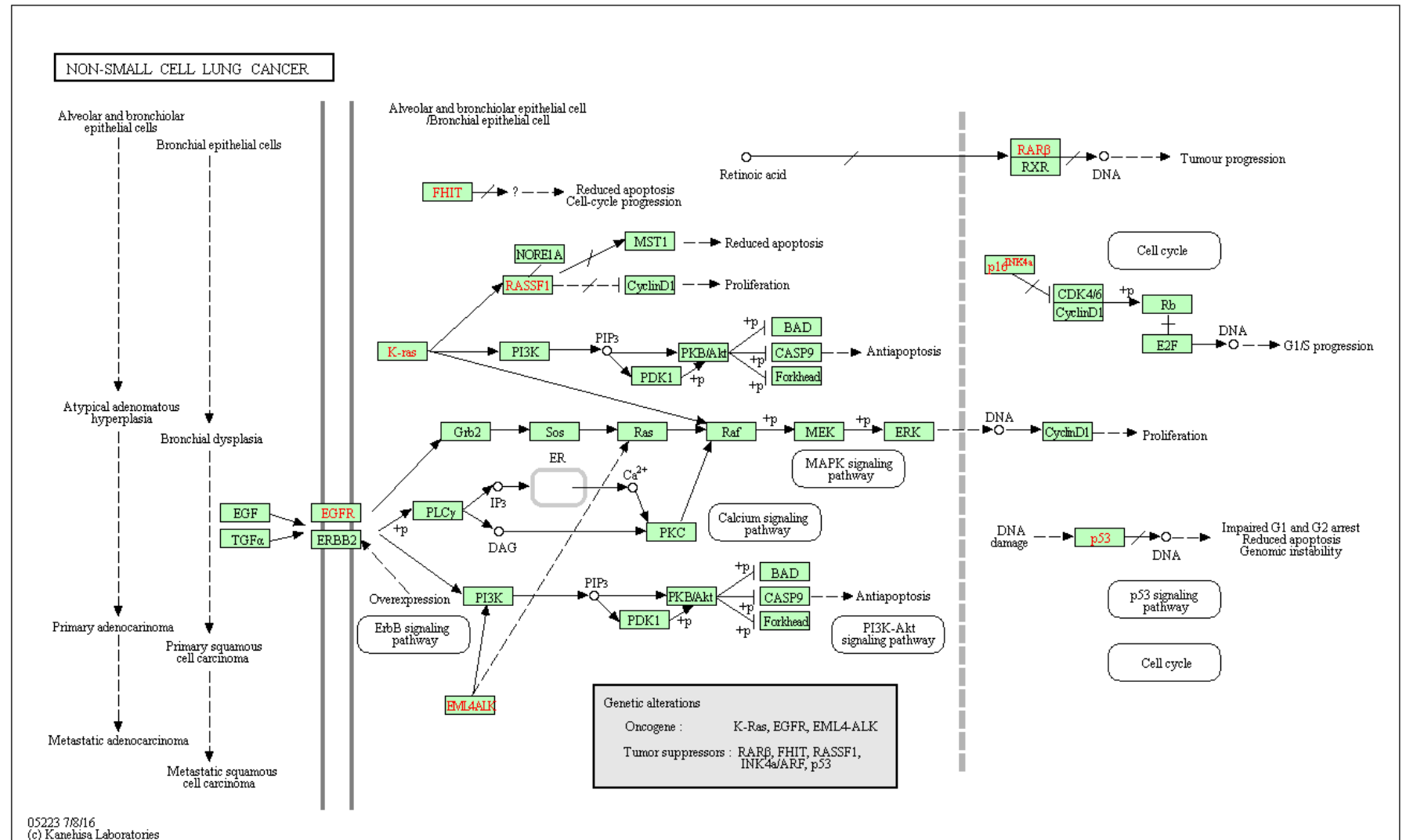
- Implication of genes in biological process
- Interaction among proteins -> Therapeutic implications
- **Resources**
 - Reactome
 - KEGG (Kyoto Encyclopedia of Genes and Genomes)

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



Wiring diagrams of molecular interactions, reactions, and relations

7. Drug development (structural relations between compounds)



Clinical Implications: ClinVar

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

	Variation Location	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
<input type="checkbox"/> 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
<input type="checkbox"/> 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
<input type="checkbox"/> 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
<input type="checkbox"/> 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
<input type="checkbox"/> 23.	NM_001005862.2(ERBB2):c.2173_2174delTTinsCC (p.Leu725Pro) GRCh37: Chr17:37880219-37880220 GRCh38: Chr17:39723966-39723967	ERBB2	Adenocarcinoma of lung		Pathogenic (Sep 30, 2004)	no assertion criteria provided

Bibliography

UniProtKB - P00533 (EGFR_HUMAN)

Protein | Epidermal growth factor receptor

Gene | EGFR

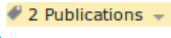






Organism | *Homo sapiens* (Human)

Status |  Reviewed - Annotation score:  - Experimental evidence at protein levelⁱ





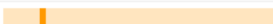

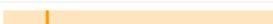


<http://www.uniprot.org>

PTM / Processingⁱ

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

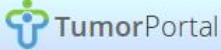
Feature key	Position(s)	Length	Description	Graphical view	Feature Identifier	Actions
Disulfide bond ⁱ	31 ↔ 58					
Glycosylation ⁱ	56 – 56	1	N-linked (GlcNAc...) (complex); atypical; partial		CAR_000227	
Glycosylation ⁱ	73 – 73	1	N-linked (GlcNAc...); atypical			
Glycosylation ⁱ	128 – 128	1	N-linked (GlcNAc...)			
Disulfide bond ⁱ	157 ↔ 187					
Glycosylation ⁱ	175 – 175	1	N-linked (GlcNAc...)			
Disulfide bond ⁱ	190 ↔ 199					
Disulfide bond ⁱ	194 ↔ 207					
Glycosylation ⁱ	196 – 196	1	N-linked (GlcNAc...)			

Specific phenotype information


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

TumorPortal

<http://www.tumorportal.org/>

 Explore dataset ▾ Annotations ▾ Contact

Sign inGo to Gene



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	<div>Other Gene</div> <div><div>Gene Symbol</div><div>Q</div></div>						

TumorPortal

Lawrence et al. Nature 2014

- 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

The screenshot displays the TumorPortal website. At the top, there is a navigation bar with the TumorPortal logo, links for 'Explore dataset', 'Annotations', and 'Contact', a 'Sign In' button, and a 'Go to Gene' search bar. Below the navigation bar is a large banner with a green silhouette of a person and the text 'Welcome to TumorPortal'. The main content area is divided into two sections. The left section, titled 'Explore dataset by tumor type', features a grid of buttons for various cancer types: AML (196 patients), BLCA (59 patients), BRCA (892 patients), MEL (118 patients), MM (207 patients), and NB (81 patients). The right section, titled 'Explore dataset by Genes', features a grid of buttons for various genes: 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 (2% of all patients), SETD2 (2% of all patients), CREBBP (2% of all patients), FBXW7 (2% of all patients), SPEN (2% of all patients), MTOR (2% of all patients), RB1 (2% of all patients), SMARCA4 (2% of all patients), and NOTCH1 (2% of all patients). A search bar for 'Gene Symbol' is located at the bottom right of the gene grid. A 'Show Annotation Activity' button is visible on the right side of the page.

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: ccl-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)



Catalogue Of Somatic Mutations In Cancer

[Projects](#) ▾ [Data](#) ▾ [Tools](#) ▾ [News](#) ▾ [Help](#) ▾ [About](#) ▾ [Genome Version](#) ▾ [SEARCH](#) [Login](#) ▾

COSMIC v83, released 07-NOV-17





COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.





[SEARCH](#)

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

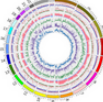
-  [COSMIC](#)
The core of COSMIC, an expert-curated database of somatic mutations
-  [Cell lines project](#)
Mutation profiles of over 1,000 cell lines used in cancer research
-  [COSMIC-3D](#)
An interactive view of cancer mutations in the context of 3D structures
-  [Cancer Gene Census](#)
A catalogue of genes with mutations that are causally implicated in cancer

Data curation

-  [Gene curation](#) — details of our manual curation process
-  [Gene fusion curation](#) — details of our curation process for gene fusions
-  [Genome Annotation](#) — Information on the annotation of genomes
-  [Drug Resistance](#) — curation of mutations conferring drug resistance


COSMIC News

[Follow @cosmic_sanger](#)




COSMIC Release v83

The November release (v83) is now live! New in this release we have 3 fully curated genes: TGFBR2, ERBB4 and BCL9L; a substantial update to VHL; and the new fusion pair ETV6-ABL1. [More...](#)



Switch between GRCh37 and GRCh38






You still have access to GRCh37 and it is now possible to switch between GRCh37 and GRCh38 from any page within the main site. [More...](#)








Cancer Gene Census changes

As we mentioned at the last release, as part of integrating the Hallmarks of Cancer feature, the Cancer Gene Census (CGC) has had a thorough re-evaluation. [More...](#)


Tools

-  [Cancer browser](#) — browse COSMIC data by tissue type and histology
-  [Genome browser](#) — browse the human genome with COSMIC annotations
-  [CONAN](#) — the COSMIC copy number analysis tool
-  [GA4GH Beacon](#) — access COSMIC data through the [GA4GH Beacon Project](#)
-  [COSMIC In BigQuery](#) — search COSMIC via the [ISB Cancer Genomics Cloud](#)


Help

-  [Downloads](#) — data that you can download from our SFTP site
-  [Documentation](#) — view our help documentation
-  [FAQ](#) — a compilation of our Frequently Asked Questions
-  [Release notes](#) — information about the latest COSMIC release
-  [Licensing](#) — information about our licensing policy

Variant / gene frequency in COSMIC



Catalogue Of Somatic Mutations In Cancer



Projects ▾Data ▾Tools ▾News ▾Help ▾About ▾Genome Version ▾ **SEARCH** **Login** ▾

COSMIC search results

Your search term "**pten**" was an exact match for the COSMIC gene [PTEN](#).

A search of the whole COSMIC database returned results in **3** sections of the database. [More...](#)

Genes (1 hit)Mutations (1968)SNPs (0)Cancer (0)Tumour Site (0)Samples (0)Pubmed (366)Studies (0)


Show entries

Gene	Alternate IDs	Tested samples	Simple Mutations	Fusions	Coding Mutations
PTEN	PTEN , ENST00000371953 , PTEN.html...	69987	3975	0	3975


Showing 1 to 1 of 1 entries

FirstPrevious1NextLast

Variant / gene frequency in COSMIC



Catalogue Of Somatic Mutations In Cancer



Projects ▾Data ▾Tools ▾News ▾Help ▾About ▾Genome Version ▾Search COSMIC...**SEARCH**Login ▾

COSMIC search results

Your search term "**pten**" was an exact match for the COSMIC gene **PTEN**.

A search of the whole COSMIC database returned results in **3** sections of the database. [More...](#)

Genes (1 hit)Mutations (1968)SNPs (0)Cancer (0)Tumour Site (0)Samples (0)Pubmed (366)Studies (0)

Show 10 ▾ entries

Gene	Syntax	Alternate IDs	Recurrence
PTEN	c.388C>G	PTEN c.388C>G...	162
PTEN	c.389G>A	PTEN c.389G>A...	134
PTEN	c.697C>T	PTEN c.697C>T...	124
PTEN	c.388C>T	PTEN c.388C>T...	94
PTEN	c.? ?del?	PTEN c.? ?del?...	91
PTEN	c.? 	PTEN c.?...	88
PTEN	c.1_1212del1212	PTEN c.1_1212del1212...	88
PTEN	c.800delA	PTEN c.800delA...	56
PTEN	c.517C>T	PTEN c.517C>T...	47
PTEN	c.950_953delTACT	PTEN c.950_953delTACT...	44

Showing 1 to 10 of 1,968 entries

FirstPrevious12345...197NextLast

Role of the gene as ONC or TSG

KRAS:

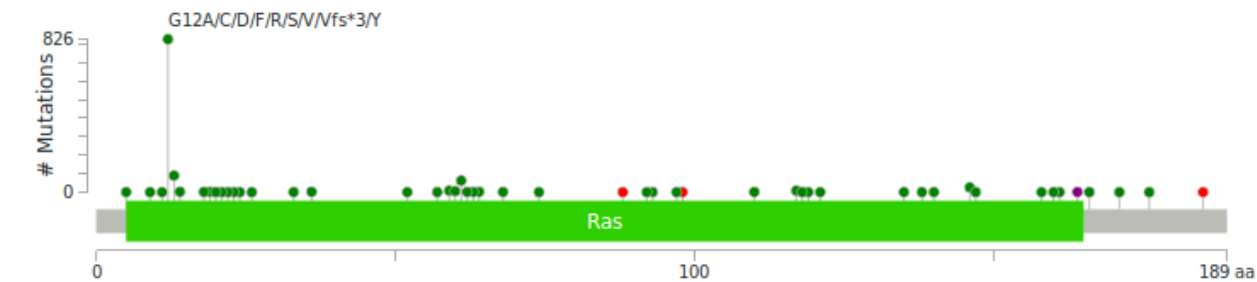
RASK_HUMAN

PDF

SVG

Customize

Color Codes



ONCOGENE

TP53:

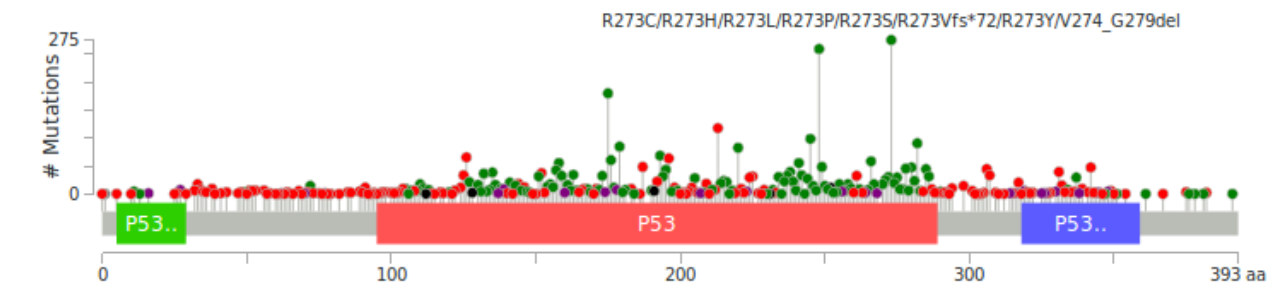
P53_HUMAN

PDF

SVG

Customize

Color Codes



TUMOR SUPPRESSOR GENE

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

Role of the gene as ONC or TSG








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

Cancer Gene Census

Showing both tiers

Show entries

Export: Search:

Gene Symbol	Name	Entrez GeneId	Genome Location	Tier	Hallmark	Chr Band	Somatic	Germline	Tumour Types(Somatic)	Tumour Types(Germline)	Cancer Syndrome	Tissue Type
ARHGAP5	Rho GTPase activating protein 5	394	14:32090670-32154948 	2		14q12	yes		colon cancer; glioma			E; O
ARHGEF12	RHO guanine nucleotide exchange factor (GEF) 12 (LARG)	23365	11:120337244-120485077 	1		11q23.3	yes		AML			L C
ARID1A	AT rich interactive domain 1A (SWI-like)	8289	1:26696404-26780756 	1		1p35.3	yes		clear cell ovarian carcinoma; RCC; breast			E R
ARID1B	AT rich interactive domain 1B	57492	6:156778104-157207891 	1		6q25.1	yes		breast; hepatocellular carcinoma			E R
ARID2	AT rich interactive domain 2	196528	12:45729837-45905078 	1		12q12	yes		hepatocellular carcinoma			E R

Role In Cancer

oncogene

TSG;
fusion

TSG;
fusion

TSG

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 0.3 1

Activating cutoff:

0 0.7 1

Cancer driver list

☐ Cancer5000

☒ HCD

[Download classification](#)

[Download & Usage](#)

[Performance and plots](#)

[Gene classification](#)

[Validation](#)

25 records per page

Search:

ENSG	SYM	oncodriveROLE	Value
ENSG00000000971	CFH	Activating	0.9180
ENSG000000009307	CSDE1	Loss of function	0.2515
ENSG000000012048	BRCA1	Loss of function	0.0020
ENSG000000019991	HGF	No class	0.4235

Filtering process

Remove artifacts

- Quality filters

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

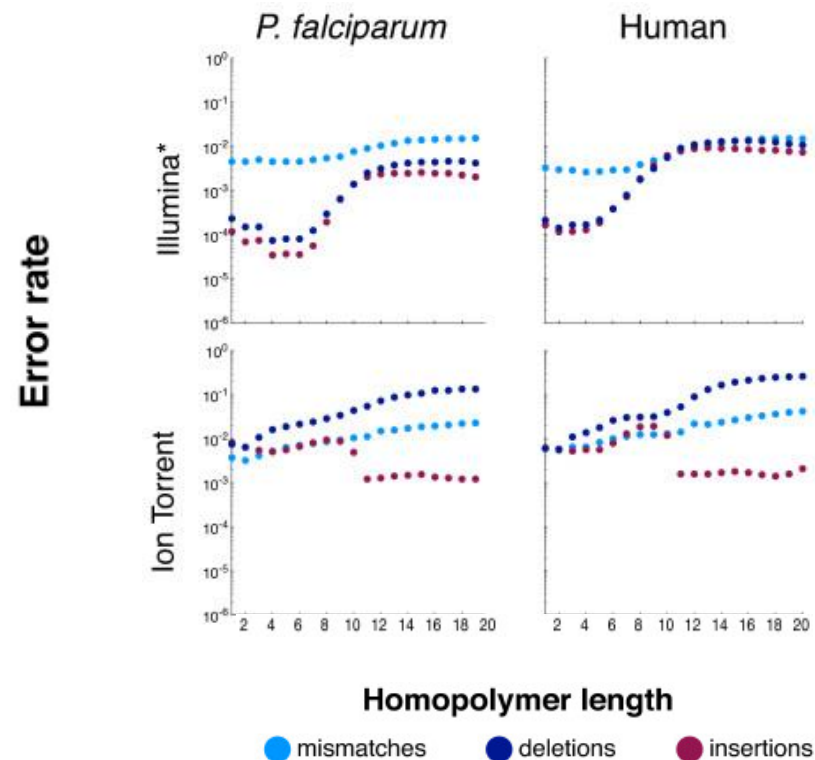
454, 454 pyrosequencing (Roche); NA, not applicable; Nanopore, Oxford Nanopore Technologies; NGS, next-generation sequencing; PacBio, Pacific Biosciences; SOLiD, sequencing by oligonucleotide ligation and detection (Thermo Fisher).

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.



Filtering process

Remove artifacts

- Quality filters
- Indels in 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 alteration if its presence is validated in other samples of the same phenotype
 - Otherwise, it can be an artifact (especially in genes acting as tumor suppressors in cancer)

Filtering process

Remove artifacts

- Quality filters
- Indels in 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 alteration if its presence is validated in other samples of the same phenotype
 - Otherwise, it can be an artifact (especially in genes acting as tumor suppressors in cancer)
- Variants in positions with very low coverage
- Variants with very low frequency

Filtering process

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

Remove variants affecting **no relevant transcripts**

Remove polymorphisms (if not interested in germline information):

Population frequency in 1000 Genomes project, ExAC, gnomAD... $\geq 1\%$

Filtering process

We can select 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

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

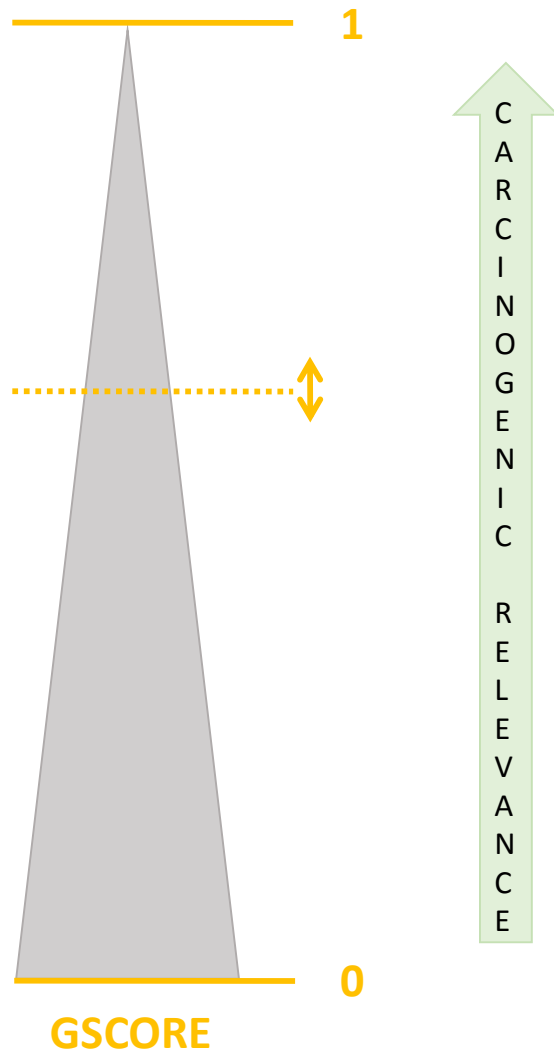
transcript_ablation | splice_donor_variant | splice_acceptor_variant | stop_gained |
frameshift_variant | stop_lost | start_lost | transcript_amplification | inframe_insertion |
inframe_deletion | missense_variant | protein_altering_variant | splice_region_variant |
incomplete_terminal_codon_variant | 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	0.125/3
Score prediction by Sift	<= 0.05	0.125/3	0.125/3
Score prediction by CONDEL	> 0.468	0.125/3	0.125/3
COSMIC	Pathogenic by FATHMM prediction	0.125/3	0.03125
Frequency of mutation in COSMIC	>= 100	0.125/3	
	< 100	$(0.125 / 3) * (\log(\text{mutation frequency}) / \log(\text{maximum mutation frequency}))$	
Frequency of gene in COSMIC	>= 100	0.125/3	0.03125
	< 100	$(0.125 / 3) * (\log(\text{gene frequency}) / \log(\text{maximum gene frequency}))$	$0.03125 * (\log(\text{gene frequency}) / \log(\text{maximum gene frequency}))$
VEP consequence	stop gain frameshift missense inframe insertion inframe deletion	0.125	0.125
GMAF	< 1	0.125/2	0.125/2
EXAC	< 1	0.125/2	0.125/2
DOMAINS	Listed as relevant in cancer [1] or previous last protein domain (in stop-gained or frameshift)	0.125	0.125
	Within a domain in other circumstances	0.125/2	0.125/2
CLINVAR	Pathogenic with zygosity data	0.125	0.125
	Pathogenic without zygosity data	0.250	0.3125
ZYGOSITY (when available)	Homozygous	0.125	0.1875
ESSENTIALITY SCORE		0.125 * ES	0.125 * ES

