

# Package ‘driveR’

November 26, 2020

**Title** Prioritizing Cancer Driver Genes Using Genomics Data

**Version** 0.2.1

**Maintainer** Ege Ulgen <egeulgen@gmail.com>

**Description** Cancer genomes contain large numbers of somatic alterations but few genes drive tumor development. Identifying cancer driver genes is critical for precision oncology. Most of current approaches either identify driver genes based on mutational recurrence or using estimated scores predicting the functional consequences of mutations. 'driveR' is a tool for personalized or batch analysis of genomic data for driver gene prioritization by combining genomic information and prior biological knowledge. As features, 'driveR' uses coding impact metaprediction scores, non-coding impact scores, somatic copy number alteration scores, hotspot gene/double-hit gene condition, 'phenolyzer' gene scores and memberships to cancer-related KEGG pathways. It uses these features to estimate cancer-type-specific probability for each gene of being a cancer driver using the related task of a multi-task learning classification model. The method is described in detail in Ulgen E, Sezerman OU. 2020. driveR: A Novel Method for Prioritizing Cancer Driver Genes Using Somatic Genomics Data. bioRxiv <doi:10.1101/2020.11.10.376707>.

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.1.1

**URL** <https://egeulgen.github.io/driveR/>,  
<https://github.com/egeulgen/driveR/>

**BugReports** <https://github.com/egeulgen/driveR/issues>

**biocViews**

**Imports** caret, randomForest, GenomicRanges, GenomeInfoDb,  
GenomicFeatures, TxDb.Hsapiens.UCSC.hg19.knownGene, S4Vectors,  
org.Hs.eg.db, rlang

**Depends** R (>= 4.0)

**Suggests** testthat, covr, knitr, rmarkdown

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Ege Ulgen [aut, cre, cph] (<<https://orcid.org/0000-0003-2090-3621>>)

**Repository** CRAN

**Date/Publication** 2020-11-26 00:30:03 UTC

## R topics documented:

create_features_df . . . . .	2
create_gene_level_scna_df . . . . .	5
create_noncoding_impact_score_df . . . . .	6
create_SCNA_score_df . . . . .	6
determine_double_hit_genes . . . . .	7
determine_hotspot_genes . . . . .	8
driveR . . . . .	9
example_cohort_features_table . . . . .	9
example_cohort_scna_table . . . . .	11
example_features_table . . . . .	11
example_scna_table . . . . .	13
KEGG_cancer_pathways . . . . .	13
KEGG_cancer_pathways_descriptions . . . . .	14
MCR_table . . . . .	14
metapredictor_model . . . . .	15
MTL_submodel_descriptions . . . . .	15
predict_coding_impact . . . . .	16
prioritize_driver_genes . . . . .	17
specific_thresholds . . . . .	19
TCGA_MTL_fit . . . . .	19
<b>Index</b>	<b>20</b>

---

create_features_df	<i>Create Data Frame of Features for Driver Gene Prioritization</i>
--------------------	---------------------------------------------------------------------

---

## Description

Create Data Frame of Features for Driver Gene Prioritization

**Usage**

```

create_features_df(
  annovar_csv_path,
  scna_df,
  phenolyzer_annotated_gene_list_path,
  batch_analysis = FALSE,
  prep_phenolyzer_input = FALSE,
  log2_ratio_threshold = 0.25,
  gene_overlap_threshold = 25,
  MCR_overlap_threshold = 25,
  hotspot_threshold = 5L,
  log2_hom_loss_threshold = -1,
  verbose = TRUE,
  na.string = "."
)

```

**Arguments**

**annovar\_csv\_path** path to 'ANNOVAR' csv output file

**scna\_df** the SCNA segments data frame. Must contain:  
**chr** chromosome the segment is located in  
**start** start position of the segment  
**end** end position of the segment  
**log2ratio**  $\log_2$  ratio of the segment

**phenolyzer\_annotated\_gene\_list\_path** path to 'phenolyzer' "annotated\_gene\_list" file

**batch\_analysis** boolean to indicate whether to perform batch analysis (TRUE, default) or personalized analysis (FALSE). If TRUE, a column named 'tumor\_id' should be present in both the ANNOVAR csv and the SCNA table.

**prep\_phenolyzer\_input** boolean to indicate whether or not to create a vector of genes for use as the input of 'phenolyzer' (default = FALSE). If TRUE, the features data frame is not created and instead the vector of gene symbols (union of all genes for which scores are available) is returned.

**log2\_ratio\_threshold** the  $\log_2$  ratio threshold for keeping high-confidence SCNA events (default = 0.25)

**gene\_overlap\_threshold** the percentage threshold for the overlap between a segment and a transcript (default = 25). This means that if only a segment overlaps a transcript more than this threshold, the transcript is assigned the segment's SCNA event.

**MCR\_overlap\_threshold** the percentage threshold for the overlap between a gene and an MCR region (default = 25). This means that if only a gene overlaps an MCR region more than this threshold, the gene is assigned the SCNA density of the MCR

hotspot_threshold	to determine hotspot genes, the (integer) threshold for the minimum number of cases with certain mutation in COSMIC (default = 5)
log2_hom_loss_threshold	to determine double-hit events, the $\log_2$ threshold for identifying homozygous loss events (default = -1).
verbose	boolean controlling verbosity (default = TRUE)
na.string	string that was used to indicate when a score is not available during annotation with ANNOVAR (default = ".")

### Value

If prep\_phenolyzer\_input=FALSE (default), a data frame of features for prioritizing cancer driver genes (gene\_symbol as the first column and 26 other columns containing features). If prep\_phenolyzer\_input=TRUE, the functions returns a vector gene symbols (union of all gene symbols for which scores are available) to be used as the input for performing 'phenolyzer' analysis.

The features data frame contains the following columns:

<b>gene_symbol</b>	HGNC gene symbol
<b>metaprediction_score</b>	the maximum metapredictor (coding) impact score for the gene
<b>noncoding_score</b>	the maximum non-coding PHRED-scaled CADD score for the gene
<b>scna_score</b>	SCNA proxy score. SCNA density (SCNA/Mb) of the minimal common region (MCR) in which the gene is located
<b>hotspot_double_hit</b>	boolean indicating whether the gene is a hotspot gene (indication of oncogenes) or subject to double-hit (indication of tumor-suppressor genes)
<b>phenolyzer_score</b>	'phenolyzer' score for the gene
<b>hsa03320</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04010</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04020</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04024</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04060</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04066</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04110</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04115</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04150</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04151</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04210</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04310</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04330</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04340</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04350</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04370</b>	boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04510** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04512** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04520** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04630** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04915** boolean indicating whether or not the gene takes part in this KEGG pathway

### See Also

[prioritize\\_driver\\_genes](#) for prioritizing cancer driver genes

### Examples

```
path2annovar_csv <- system.file("extdata/example.hg19_multianno.csv",
                               package = "driveR")
path2phenolyzer_out <- system.file("extdata/example.annotated_gene_list",
                                   package = "driveR")
features_df <- create_features_df(annovar_csv_path = path2annovar_csv,
                                 scna_df = example_scna_table,
                                 phenolyzer_annotated_gene_list_path = path2phenolyzer_out)
```

---

create\_gene\_level\_scna\_df

*Create Gene-level SCNA Data Frame*

---

### Description

Create Gene-level SCNA Data Frame

### Usage

```
create_gene_level_scna_df(scna_df, gene_overlap_threshold = 25)
```

### Arguments

**scna\_df** the SCNA segments data frame. Must contain:

**chr** chromosome the segment is located in

**start** start position of the segment

**end** end position of the segment

**log2ratio**  $\log_2$  ratio of the segment

**gene\_overlap\_threshold**

the percentage threshold for the overlap between a segment and a transcript (default = 25). This means that if only a segment overlaps a transcript more than this threshold, the transcript is assigned the segment's SCNA event.

**Value**

data frame of gene-level SCNA events, i.e. table of genes overlapped by SCNA segments.

---

```
create_noncoding_impact_score_df
```

*Create Non-coding Impact Score Data Frame*

---

**Description**

Create Non-coding Impact Score Data Frame

**Usage**

```
create_noncoding_impact_score_df(annovar_csv_path, na.string = ".")
```

**Arguments**

annovar_csv_path	path to 'ANNOVAR' csv output file
na.string	string that was used to indicate when a score is not available during annotation with ANNOVAR (default = ".")

**Value**

data frame of meta-prediction scores containing 2 columns:

**gene\_symbol** HGNC gene symbol  
**CADD\_phred** PHRED-scaled CADD score

---

```
create_SCNA_score_df
```

*Create SCNA Score Data Frame*

---

**Description**

Create SCNA Score Data Frame

**Usage**

```
create_SCNA_score_df(  
  gene_SCNA_df,  
  log2_ratio_threshold = 0.25,  
  MCR_overlap_threshold = 25  
)
```

### Arguments

**gene\_SCNA\_df** data frame of gene-level SCNAs (output of `create_gene_level_scna_df`)

**log2\_ratio\_threshold**  
the  $\log_2$  ratio threshold for keeping high-confidence SCNA events (default = 0.25)

**MCR\_overlap\_threshold**  
the percentage threshold for the overlap between a gene and an MCR region (default = 25). This means that if only a gene overlaps an MCR region more than this threshold, the gene is assigned the SCNA density of the MCR

### Details

The function first aggregates SCNA  $\log_2$  ratio on gene-level (by keeping the ratio with the maximal  $|\log_2|$  ratio over all the SCNA segments overlapping a gene). Next, it identifies the minimal common regions (MCRs) that the genes overlap and finally assigns the SCNA density (SCNA/Mb) values as proxy SCNA scores.

### Value

data frame of SCNA proxy scores containing 2 columns:

**gene\_symbol** HGNC gene symbol

**SCNA\_density** SCNA proxy score. SCNA density (SCNA/Mb) of the minimal common region (MCR) in which the gene is located.

---

determine\_double\_hit\_genes

*Determine Double-Hit Genes*

---

### Description

Determine Double-Hit Genes

### Usage

```
determine_double_hit_genes(  
  annovar_csv_path,  
  gene_SCNA_df,  
  log2_hom_loss_threshold = -1,  
  batch_analysis = FALSE  
)
```

**Arguments**

- `annovar_csv_path` path to 'ANNOVAR' csv output file
- `gene_SCNA_df` data frame of gene-level SCNAs (output of `create_gene_level_scna_df`)
- `log2_hom_loss_threshold` to determine double-hit events, the  $\log_2$  threshold for identifying homozygous loss events (default = -1).
- `batch_analysis` boolean to indicate whether to perform batch analysis (TRUE, default) or personalized analysis (FALSE). If TRUE, a column named 'tumor\_id' should be present in both the ANNOVAR csv and the SCNA table.

**Value**

vector of gene symbols that are subject to double-hit event(s), i.e. non-synonymous mutation + homozygous copy-number loss

---

`determine_hotspot_genes`

*Determine Hotspot Containing Genes*

---

**Description**

Determine Hotspot Containing Genes

**Usage**

```
determine_hotspot_genes(annovar_csv_path, hotspot_threshold = 5L)
```

**Arguments**

- `annovar_csv_path` path to 'ANNOVAR' csv output file
- `hotspot_threshold` to determine hotspot genes, the (integer) threshold for the minimum number of cases with certain mutation in COSMIC (default = 5)

**Value**

vector of gene symbols of genes containing hotspot mutation(s)

---

driveR	<i>driveR: An R Package for Prioritizing Cancer Driver Genes Using Genomics Data</i>
--------	--------------------------------------------------------------------------------------

---

## Description

Cancer genomes contain large numbers of somatic alterations but few genes drive tumor development. Identifying cancer driver genes is critical for precision oncology. Most of current approaches either identify driver genes based on mutational recurrence or using estimated scores predicting the functional consequences of mutations.

## Details

driveR is a tool for personalized or batch analysis of genomic data for driver gene prioritization by combining genomic information and prior biological knowledge. As features, driveR uses coding impact metaprediction scores, non-coding impact scores, somatic copy number alteration scores, hotspot gene/double-hit gene condition, 'phenolyzer' gene scores and memberships to cancer-related KEGG pathways. It uses these features to estimate cancer-type-specific probabilities for each gene of being a cancer driver using the related task of a multi-task learning classification model.

## See Also

[predict\\_coding\\_impact](#) for metaprediction of impact of coding variants. [create\\_features\\_df](#) for creating the features table to be used to prioritize cancer driver genes. See [prioritize\\_driver\\_genes](#) for prioritizing cancer driver genes

---

example_cohort_features_table	<i>Example Cohort-level Features Table for Driver Prioritization</i>
-------------------------------	----------------------------------------------------------------------

---

## Description

The example dataset containing features for prioritizing cancer driver genes for 10 randomly selected samples from TCGA's LAML (Acute Myeloid Leukemia) cohort

## Usage

example\_cohort\_features\_table

**Format**

A data frame with 349 rows and 27 variables:

**gene\_symbol** HGNC gene symbol

**metaprediction\_score** the maximum metapredictor (coding) impact score for the gene

**noncoding\_score** the maximum non-coding PHRED-scaled CADD score for the gene

**scna\_score** SCNA proxy score. SCNA density (SCNA/Mb) of the minimal common region (MCR) in which the gene is located

**hotspot\_double\_hit** boolean indicating whether the gene is a hotspot gene (indication of oncogenes) or subject to double-hit (indication of tumor-suppressor genes)

**phenolyzer\_score** 'phenolyzer' score for the gene

**hsa03320** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04010** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04020** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04024** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04060** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04066** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04110** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04115** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04150** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04151** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04210** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04310** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04330** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04340** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04350** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04370** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04510** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04512** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04520** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04630** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04915** boolean indicating whether or not the gene takes part in this KEGG pathway

**See Also**

[KEGG\\_cancer\\_pathways\\_descriptions](#) for descriptions of KEGG "Pathways in cancer"-related pathways.

---

`example_cohort_scna_table`*Example Cohort-level Somatic Copy Number Alteration Table*

---

**Description**

A data set containing the somatic copy number alteration data for 10 randomly selected samples from TCGA's LAML (Acute Myeloid Leukemia) cohort

**Usage**`example_cohort_scna_table`**Format**

A data frame with 126147 rows and 5 variables:

**chr** chromosome the segment is located in

**start** start position of the segment

**end** end position of the segment

**log2ratio**  $\log_2$  ratio of the segment

**tumor\_id** ID for the tumor containing the SCNA segment

**Source**

[https://dcc.icgc.org/releases/release\\_28](https://dcc.icgc.org/releases/release_28)

---

`example_features_table`*Example Features Table for Driver Prioritization*

---

**Description**

The example dataset containing features for prioritizing cancer driver genes for the lung adenocarcinoma patient studied in Imielinski M, Greulich H, Kaplan B, et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. J Clin Invest. 2014;124(4):1582-6.

**Usage**`example_features_table`

**Format**

A data frame with 4901 rows and 27 variables:

**gene\_symbol** HGNC gene symbol

**metaprediction\_score** the maximum metapredictor (coding) impact score for the gene

**noncoding\_score** the maximum non-coding PHRED-scaled CADD score for the gene

**scna\_score** SCNA proxy score. SCNA density (SCNA/Mb) of the minimal common region (MCR) in which the gene is located

**hotspot\_double\_hit** boolean indicating whether the gene is a hotspot gene (indication of oncogenes) or subject to double-hit (indication of tumor-suppressor genes)

**phenolyzer\_score** 'phenolyzer' score for the gene

**hsa03320** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04010** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04020** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04024** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04060** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04066** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04110** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04115** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04150** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04151** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04210** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04310** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04330** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04340** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04350** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04370** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04510** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04512** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04520** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04630** boolean indicating whether or not the gene takes part in this KEGG pathway

**hsa04915** boolean indicating whether or not the gene takes part in this KEGG pathway

**See Also**

[KEGG\\_cancer\\_pathways\\_descriptions](#) for descriptions of KEGG "Pathways in cancer"-related pathways.

---

example_scna_table	<i>Example Somatic Copy Number Alteration Table</i>
--------------------	-----------------------------------------------------

---

**Description**

A data set containing the somatic copy number alteration data for the lung adenocarcinoma patient studied in Imielinski M, Greulich H, Kaplan B, et al. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. J Clin Invest. 2014;124(4):1582-6.

**Usage**

```
example_scna_table
```

**Format**

A data frame with 3160 rows and 4 variables:

**chr** chromosome the segment is located in

**start** start position of the segment

**end** end position of the segment

**log2ratio**  $\log_2$  ratio of the segment

**Source**

<https://pubmed.ncbi.nlm.nih.gov/24569458/>

---

KEGG_cancer_pathways	<i>KEGG "Pathways in cancer"-related Pathways - Gene Sets</i>
----------------------	---------------------------------------------------------------

---

**Description**

A list containing the genes involved in each Homo sapiens KEGG "Pathways in cancer" (hsa05200)-related Pathways. Each element is a vector of gene symbols located in the given pathway. Names correspond to the KEGG ID of the pathway. *Generated on Nov 24, 2020.*

**Usage**

```
KEGG_cancer_pathways
```

**Format**

list containing 21 vectors of gene symbols. Each vector corresponds to a pathway.

**See Also**

[KEGG\\_cancer\\_pathways\\_descriptions](#) for descriptions of KEGG "Pathways in cancer"-related pathways.

---

KEGG\_cancer\_pathways\_descriptions

*KEGG "Pathways in cancer"-related Pathways - Descriptions*

---

### Description

A data frame containing descriptions for KEGG "Pathways in cancer" (hsa05200)-related pathways.  
Generated on Nov 17, 2020.

### Usage

KEGG\_cancer\_pathways\_descriptions

### Format

A data frame with 21 rows and 2 variables:

**id** KEGG pathway ID

**description** KEGG pathway description

---

MCR\_table

*Table of Pan-Cancer Minimal Common Regions*

---

### Description

A data set containing the minimal common regions (MCRs) across all cancer types studied in Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* 2013;23(2):217-27. Coordinates were converted to hg19 (from hg18) using UCSC Genome Browser's LiftOver tool.

### Usage

MCR\_table

### Format

A data frame with 165 rows and 5 variables:

**chr** chromosome the MCR is located in

**start** start position of the MCR

**end** end position of the MCR

**MCR\_type** the type ("Amp" or "Del") of the MCR peak

**SCNA\_density** SCNA per Mb within the MCR

### Source

<https://pubmed.ncbi.nlm.nih.gov/23132910/>

---

metapredictor\_model     *Random Forest Model for Coding Impact Metaprediction*

---

**Description**

A Random Forest model object for metaprediction of coding variants' impact, using 12 impact scores from different coding impact predictors. The model was trained on 711 coding variants, with 10-folds repeated 3 times cross-validation.

**Usage**

metapredictor\_model

**Format**

model object

---

MTL\_submodel\_descriptions  
*MTL Sub-model Descriptions*

---

**Description**

A data frame containing descriptions for all sub-models of the MTL model.

**Usage**

MTL\_submodel\_descriptions

**Format**

A data frame with 21 rows and 2 variables:

**short\_name** short name for the cancer type

**description** description of the cancer type

**See Also**

[TCGA\\_MTL\\_fit](#) for the MTL model.

---

predict\_coding\_impact *Create Coding Impact Meta-prediction Score Data Frame*

---

## Description

Create Coding Impact Meta-prediction Score Data Frame

## Usage

```
predict_coding_impact(  
  annovar_csv_path,  
  keep_highest_score = TRUE,  
  keep_single_symbol = TRUE,  
  na.string = ".")  
)
```

## Arguments

**annovar\_csv\_path** path to 'ANNOVAR' csv output file

**keep\_highest\_score** boolean to indicate whether to keep only the maximal impact score per gene (default = TRUE). If FALSE, all scores per each gene are returned

**keep\_single\_symbol** in ANNOVAR outputs, a variant may be annotated as exonic in multiple genes. This boolean argument controls whether or not to keep only the first encountered symbol for a variant (default = TRUE)

**na.string** string that was used to indicate when a score is not available during annotation with ANNOVAR (default = ".")

## Value

data frame of meta-prediction scores containing 2 columns:

**gene\_symbol** HGNC gene symbol

**metaprediction\_score** metapredictor impact score

## Examples

```
path2annovar_csv <- system.file("extdata/example.hg19_multianno.csv",  
                                package = "driveR")  
metapred_df <- predict_coding_impact(path2annovar_csv)
```

---

`prioritize_driver_genes`*Prioritize Cancer Driver Genes*

---

## Description

Prioritize Cancer Driver Genes

## Usage

```
prioritize_driver_genes(features_df, cancer_type)
```

## Arguments

`features_df` the features data frame for all genes, containing the following columns:

- gene\_symbol** HGNC gene symbol
- metaprediction\_score** the maximum metapredictor (coding) impact score for the gene
- noncoding\_score** the maximum non-coding PHRED-scaled CADD score for the gene
- scna\_score** SCNA proxy score. SCNA density (SCNA/Mb) of the minimal common region (MCR) in which the gene is located
- hotspot\_double\_hit** boolean indicating whether the gene is a hotspot gene (indication of oncogenes) or subject to double-hit (indication of tumor-suppressor genes)
- phenolyzer\_score** 'phenolyzer' score for the gene
- hsa03320** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04010** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04020** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04024** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04060** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04066** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04110** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04115** boolean indicating whether or not the gene takes part in this KEGG pathway
- hsa04150** boolean indicating whether or not the gene takes part in this KEGG pathway

<b>hsa04151</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04210</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04310</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04330</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04340</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04350</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04370</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04510</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04512</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04520</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04630</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
<b>hsa04915</b>	boolean indicating whether or not the gene takes part in this KEGG pathway
cancer_type	short name of the cancer type. All available cancer types are listed in <a href="#">MTL_submodel_descriptions</a>

**Value**

data frame with 3 columns:

**gene\_symbol** HGNC gene symbol

**driverness\_prob** estimated probability for each gene in `features_df` of being a cancer driver. The probabilities are calculated using the selected (via `cancer_type`) cancer type's sub-model.

**prediction** prediction based on the cancer-type-specific threshold (either "driver" or "non-driver")

**See Also**

[create\\_features\\_df](#) for creating the features table. [TCGA\\_MTL\\_fit](#) for details on the MTL model used for prediction.

**Examples**

```
drivers_df <- prioritize_driver_genes(example_features_table, "LUAD")
```

---

specific_thresholds	<i>Tumor type specific probability thresholds</i>
---------------------	---------------------------------------------------

---

**Description**

Driver gene probability thresholds for all 21 cancer types (submodels).

**Usage**

specific\_thresholds

**Format**

vector with 21 elements

**See Also**

[TCGA\\_MTL\\_fit](#) for the Multi-Task Learning model.

---

TCGA_MTL_fit	<i>Multi-Task Learning Model for Predicting Cancer Driver Genes</i>
--------------	---------------------------------------------------------------------

---

**Description**

A Multi-Task Learning (MTL) classification model object for determining cancer driver genes based on 26 features. The model was trained using TCGA data (obtained from ICGC release 28) with lasso regularization. It contains 21 sub-models for different cancer types.

**Usage**

TCGA\_MTL\_fit

**Format**

MTL model object

**See Also**

[MTL\\_submodel\\_descriptions](#) for short names and descriptions of all sub-models.

# Index

## \* datasets

- example\_cohort\_features\_table, 9
- example\_cohort\_scna\_table, 11
- example\_features\_table, 11
- example\_scna\_table, 13
- KEGG\_cancer\_pathways, 13
- KEGG\_cancer\_pathways\_descriptions, 14
- MCR\_table, 14
- metapredictor\_model, 15
- MTL\_submodel\_descriptions, 15
- specific\_thresholds, 19
- TCGA\_MTL\_fit, 19

- create\_features\_df, 2, 9, 18
- create\_gene\_level\_scna\_df, 5, 7, 8
- create\_noncoding\_impact\_score\_df, 6
- create\_SCNA\_score\_df, 6

- determine\_double\_hit\_genes, 7
- determine\_hotspot\_genes, 8
- driverR, 9

- example\_cohort\_features\_table, 9
- example\_cohort\_scna\_table, 11
- example\_features\_table, 11
- example\_scna\_table, 13

- KEGG\_cancer\_pathways, 13
- KEGG\_cancer\_pathways\_descriptions, 10, 12, 13, 14

- MCR\_table, 14
- metapredictor\_model, 15
- MTL\_submodel\_descriptions, 15, 18, 19

- predict\_coding\_impact, 9, 16
- prioritize\_driver\_genes, 5, 9, 17

- specific\_thresholds, 19

- TCGA\_MTL\_fit, 15, 18, 19, 19