

# COSMIC 67

*Julian Gehring, EMBL Heidelberg*

October 26, 2023

## Contents

1	Introduction . . . . .	1
2	Accessing and Using the Data . . . . .	1
3	Data Provenance . . . . .	4
3.1	COSMIC Mutations . . . . .	4
3.2	Cancer Gene Census . . . . .	4
4	Data Source . . . . .	5
5	References . . . . .	5
6	Session Info . . . . .	5

## 1 Introduction

---

The *COSMIC.67* package provides the curated mutations published with the COSMIC release version 67 (2013-10-24). Both variants found in coding and non-coding regions are included and offered as a single object of class 'CollapsedVCF' and a bgzipped and tabix-index 'VCF' file.

Additionally, the package contains the Cancer Gene Census, a list of genes causally linked to cancer.

## 2 Accessing and Using the Data

---

```
library(VariantAnnotation)
```

```
Loading required package: BiocGenerics
```

```
Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:stats':
```

```
  IQR, mad, sd, var, xtabs
```

```
The following objects are masked from 'package:base':
```

## COSMIC 67

*Filter, Find, Map, Position, Reduce, anyDuplicated, aperm, append, as.data.frame, basename, cbind, colnames, dirname, do.call, duplicated, eval, evalq, get, grep, grepl, intersect, is.unsorted, lapply, mapply, match, mget, order, paste, pmax, pmax.int, pmin, pmin.int, rank, rbind, rownames, sapply, setdiff, sort, table, tapply, union, unique, unsplit, which.max, which.min*

Loading required package: *MatrixGenerics*

Loading required package: *matrixStats*

Attaching package: *'MatrixGenerics'*

The following objects are masked from *'package:matrixStats'*:

*colAlls, colAnyNAs, colAnys, colAvgPerRowSet, colCollapse, colCounts, colCummaxs, colCummins, colCumprods, colCumsums, colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs, colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats, colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds, colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads, colWeightedMeans, colWeightedMedians, colWeightedSds, colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgPerColSet, rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods, rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps, rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins, rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks, rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars, rowWeightedMads, rowWeightedMeans, rowWeightedMedians, rowWeightedSds, rowWeightedVars*

Loading required package: *GenomeInfoDb*

Loading required package: *S4Vectors*

Loading required package: *stats4*

Attaching package: *'S4Vectors'*

The following object is masked from *'package:utils'*:

*findMatches*

The following objects are masked from *'package:base'*:

*I, expand.grid, unname*

Loading required package: *IRanges*

Loading required package: *GenomicRanges*

## COSMIC 67

```
Loading required package: SummarizedExperiment
Loading required package: Biobase
Welcome to Bioconductor

  Vignettes contain introductory material; view with
  'browseVignettes()'. To cite Bioconductor, see
  'citation("Biobase)", and for packages
  'citation("pkgname)".

Attaching package: 'Biobase'

The following object is masked from 'package:MatrixGenerics':
  rowMedians

The following objects are masked from 'package:matrixStats':
  anyMissing, rowMedians

Loading required package: Rsamtools
Loading required package: Biostrings
Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:base':
  strsplit

Attaching package: 'VariantAnnotation'

The following object is masked from 'package:base':
  tabulate

library(GenomicRanges)

data(package = "COSMIC.67")
data(cosmic_67, package = "COSMIC.67")

tp53_range = GRanges("17", IRanges(7565097, 7590856))
vcf_path = system.file("vcf", "cosmic_67.vcf.gz", package = "COSMIC.67")
cosmic_tp53 = readVcf(vcf_path, genome = "GRCh37", ScanVcfParam(which = tp53_range))
cosmic_tp53

class: CollapsedVCF
dim: 5892 0
rowRanges(vcf):
  GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
info(vcf):
  DataFrame with 5 columns: GENE, STRAND, CDS, AA, CNT
info(header(vcf)):
      Number Type      Description
```

## COSMIC 67

```
GENE 1 String Gene name
STRAND 1 String Gene strand
CDS 1 String CDS annotation
AA 1 String Peptide annotation
CNT 1 Integer How many samples have this mutation
geno(vcf):
List of length 0:
```

```
data(cgc_67, package = "COSMIC.67")
head(cgc_67)
```

```
SYMBOL ENTREZID ENSEMBL
1 ABI1 10006 ENSG00000136754
2 ABL1 25 ENSG00000097007
3 ABL2 27 ENSG00000143322
4 ACSL3 2181 ENSG00000123983
5 CASC5 57082 ENSG00000137812
6 MLLT11 10962 ENSG00000213190
```

For details on the collection and curation of the original data, please see the webpage of the COSMIC project: <http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/>.

## 3 Data Provenance

---

### 3.1 COSMIC Mutations

The following steps are performed for importing and processing of the VCF data:

1. Downloading of the VCF files 'CosmicCodingMuts\_v67\_20131024.vcf.gz' and 'Cosmic-NonCodingVariants\_v67\_20131024.vcf.gz' from 'ftp://ngs.sanger.ac.uk/production/cosmic/' to 'inst/raw/'.
2. Importing of both files to R using 'readVcf'.
3. Sorting of the seqlevels and adding 'seqinfo' data for the toplevel chromosomes of 'GRCh37'.
4. Merging of both objects, sorting according to genomic position.
5. Converting the object to class `VariantAnnotation::VRanges`.
6. Converting the 'character' columns to 'factors'.
7. Saving the merged object to 'data/cosmic\_v67\_vcf.rda'.
8. Exporting the merged object as a bgzipped and tabix-indexed 'VCF' to 'inst/vcf/cosmic\_v67.vcf.gz'.

### 3.2 Cancer Gene Census

The following steps are performed for importing and processing of the Cancer Gene Census data:

1. Downloading of the 'cancer\_gene\_census.tsv' file from [ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/data\\_export](ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/data_export) to 'inst/raw'.

2. Import of the files as a data frame.
3. Annotation of the 'HGNC' and 'ENSEMBLID' identifiers, using the 'ENTREZ gene ID' as query with the 'org.Hs.eg.db' object.
4. Saving the object to 'data/cgc\_67.rda'.

## 4 Data Source

---

The mutation data was obtained from the Sanger Institute Catalogue Of Somatic Mutations In Cancer web site, <http://www.sanger.ac.uk/cosmic>

Bamford et al (2004):

The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.

Br J Cancer, 91,355-358.

For details on the usage and redistribution of the data, please see [ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/GUIDELINES\\_ON\\_THE\\_USE\\_OF\\_THIS\\_DATA.txt](ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/GUIDELINES_ON_THE_USE_OF_THIS_DATA.txt).

## 5 References

---

- <http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/>
- [http://nar.oxfordjournals.org/content/39/suppl\\_1/D945.long](http://nar.oxfordjournals.org/content/39/suppl_1/D945.long)
- [ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/GUIDELINES\\_ON\\_THE\\_USE\\_OF\\_THIS\\_DATA.txt](ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/GUIDELINES_ON_THE_USE_OF_THIS_DATA.txt)

## 6 Session Info

---

```
R version 4.3.1 (2023-06-16)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 22.04.3 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.18-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
```

```
locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
 [3] LC_TIME=en_GB            LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
 [9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
time zone: America/New_York
tzcode source: system (glibc)
```

```
attached base packages:
[1] stats4      stats      graphics  grDevices  utils      datasets
```

## COSMIC 67

[7] methods base

other attached packages:

[1] VariantAnnotation_1.48.0	Rsamtools_2.18.0
[3] Biostrings_2.70.1	XVector_0.42.0
[5] SummarizedExperiment_1.32.0	Biobase_2.62.0
[7] GenomicRanges_1.54.0	GenomeInfoDb_1.38.0
[9] IRanges_2.36.0	S4Vectors_0.40.0
[11] MatrixGenerics_1.14.0	matrixStats_1.0.0
[13] BiocGenerics_0.48.0	knitr_1.44

loaded via a namespace (and not attached):

[1] tidyselect_1.2.0	dplyr_1.1.3
[3] blob_1.2.4	filelock_1.0.2
[5] bitops_1.0-7	fastmap_1.1.1
[7] RCurl_1.98-1.12	BiocFileCache_2.10.0
[9] GenomicAlignments_1.38.0	XML_3.99-0.14
[11] digest_0.6.33	lifecycle_1.0.3
[13] KEGGREST_1.42.0	RSQLite_2.3.1
[15] magrittr_2.0.3	compiler_4.3.1
[17] rlang_1.1.1	progress_1.2.2
[19] tools_4.3.1	utf8_1.2.4
[21] yaml_2.3.7	rtracklayer_1.62.0
[23] prettyunits_1.2.0	S4Arrays_1.2.0
[25] bit_4.0.5	curl_5.1.0
[27] DelayedArray_0.28.0	xml2_1.3.5
[29] abind_1.4-5	BiocParallel_1.36.0
[31] grid_4.3.1	fansi_1.0.5
[33] biomaRt_2.58.0	cli_3.6.1
[35] rmarkdown_2.25	crayon_1.5.2
[37] generics_0.1.3	httr_1.4.7
[39] rjson_0.2.21	DBI_1.1.3
[41] cachem_1.0.8	stringr_1.5.0
[43] zlibbioc_1.48.0	parallel_4.3.1
[45] AnnotationDbi_1.64.0	BiocManager_1.30.22
[47] restfulr_0.0.15	vctrs_0.6.4
[49] Matrix_1.6-1.1	hms_1.1.3
[51] bit64_4.0.5	GenomicFeatures_1.54.0
[53] glue_1.6.2	codetools_0.2-19
[55] stringi_1.7.12	BiocIO_1.12.0
[57] tibble_3.2.1	pillar_1.9.0
[59] rappdirs_0.3.3	htmltools_0.5.6.1
[61] GenomeInfoDbData_1.2.11	BSgenome_1.70.0
[63] R6_2.5.1	dbplyr_2.3.4
[65] evaluate_0.22	lattice_0.22-5
[67] highr_0.10	png_0.1-8
[69] memoise_2.0.1	BiocStyle_2.30.0
[71] SparseArray_1.2.0	xfun_0.40
[73] pkgconfig_2.0.3	