

COSMIC 67

Julian Gehring, EMBL Heidelberg

April 27, 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

Warning: replacing previous import 'utils::findMatches' by 'S4Vectors::findMatches'
when loading 'AnnotationDbi'

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):
```

COSMIC 67

```
DataFrame with 5 columns: GENE, STRAND, CDS, AA, CNT
info(header(vcf)):
  Number Type      Description
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 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.

5 References

- <http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/>
- 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

6 Session Info

R version 4.3.0 RC (2023-04-18 r84287)

Platform: x86_64-pc-linux-gnu (64-bit)

Running under: Ubuntu 22.04.2 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)

COSMIC 67

attached base packages:

```
[1] stats4      stats      graphics  grDevices  utils      datasets
[7] methods    base
```

other attached packages:

```
[1] VariantAnnotation_1.47.0   Rsamtools_2.17.0
[3] Biostrings_2.69.0         XVector_0.41.0
[5] SummarizedExperiment_1.31.0 Biobase_2.61.0
[7] GenomicRanges_1.53.0      GenomeInfoDb_1.37.0
[9] IRanges_2.35.0            S4Vectors_0.39.0
[11] MatrixGenerics_1.13.0     matrixStats_0.63.0
[13] BiocGenerics_0.47.0       knitr_1.42
```

loaded via a namespace (and not attached):

```
[1] KEGGREST_1.41.0          rjson_0.2.21
[3] xfun_0.39                lattice_0.21-8
[5] vctrs_0.6.2              tools_4.3.0
[7] bitops_1.0-7             generics_0.1.3
[9] curl_5.0.0               parallel_4.3.0
[11] tibble_3.2.1             fansi_1.0.4
[13] AnnotationDbi_1.63.0     RSQLite_2.3.1
[15] highr_0.10               blob_1.2.4
[17] pkgconfig_2.0.3          Matrix_1.5-4
[19] BSgenome_1.69.0          dbplyr_2.3.2
[21] lifecycle_1.0.3          GenomeInfoDbData_1.2.10
[23] compiler_4.3.0           stringr_1.5.0
[25] progress_1.2.2           BiocStyle_2.29.0
[27] codetools_0.2-19         htmltools_0.5.5
[29] RCurl_1.98-1.12         yaml_2.3.7
[31] pillar_1.9.0             crayon_1.5.2
[33] BiocParallel_1.35.0      DelayedArray_0.27.0
[35] cachem_1.0.7             tidyselect_1.2.0
[37] digest_0.6.31           stringi_1.7.12
[39] restfulr_0.0.15         dplyr_1.1.2
[41] biomaRt_2.57.0          fastmap_1.1.1
[43] grid_4.3.0              cli_3.6.1
[45] magrittr_2.0.3          GenomicFeatures_1.53.0
[47] utf8_1.2.3              XML_3.99-0.14
[49] rappdirs_0.3.3          filelock_1.0.2
[51] prettyunits_1.1.1       bit64_4.0.5
[53] rmarkdown_2.21          httr_1.4.5
[55] bit_4.0.5               png_0.1-8
[57] hms_1.1.3               memoise_2.0.1
[59] evaluate_0.20           BiocIO_1.11.0
[61] BiocFileCache_2.9.0     rtracklayer_1.61.0
[63] rlang_1.1.0             glue_1.6.2
[65] DBI_1.1.3               xml2_1.3.3
[67] BiocManager_1.30.20     R6_2.5.1
[69] GenomicAlignments_1.37.0 zlibbioc_1.47.0
```