COSMIC 67

Julian Gehring, EMBL Heidelberg

November 5, 2024

Contents

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
```

```
Loading required package:  generics
```

```

Attaching package:  'generics'
```

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

```
   as.difftime, as.factor, as.ordered, intersect,
is.element, setdiff, setequal, union


Attaching package:  'BiocGenerics'

The following objects are masked from 'package:generics':

   intersect, setdiff, setequal, union

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

   IQR, mad, sd, var, xtabs

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

   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,
saveRDS, setdiff, setequal, 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, colAvgsPerRowSet,
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,
rowAvgsPerColSet, 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

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
   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 'CosmicNonCodingVariants_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.oxfordjournalls.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 Under development (unstable) (2024-10-21 r87258)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.1 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.21-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0

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
[7] methods   base

other attached packages:
 [1] VariantAnnotation_1.53.0    Rsamtools_2.23.0
 [3] Biostrings_2.75.0           XVector_0.47.0
 [5] SummarizedExperiment_1.37.0 Biobase_2.67.0
 [7] GenomicRanges_1.59.0        GenomeInfoDb_1.43.0
 [9] IRanges_2.41.0              S4Vectors_0.45.0
[11] MatrixGenerics_1.19.0       matrixStats_1.4.1
[13] BiocGenerics_0.53.1         generics_0.1.3
[15] knitr_1.48

loaded via a namespace (and not attached):
 [1] SparseArray_1.7.0        bitops_1.0-9
 [3] RSQLite_2.3.7            lattice_0.22-6
 [5] digest_0.6.37            evaluate_1.0.1
 [7] grid_4.5.0               fastmap_1.2.0
 [9] blob_1.2.4               jsonlite_1.8.9
[11] Matrix_1.7-1             AnnotationDbi_1.69.0
[13] restfulr_0.0.15          DBI_1.2.3
[15] BiocManager_1.30.25      httr_1.4.7
[17] BSgenome_1.75.0          UCSC.utils_1.3.0
[19] XML_3.99-0.17            codetools_0.2-20
[21] abind_1.4-8              cli_3.6.3
[23] rlang_1.1.4              crayon_1.5.3
[25] BiocStyle_2.35.0         bit64_4.5.2
[27] cachem_1.1.0             DelayedArray_0.33.1
[29] yaml_2.3.10              GenomicFeatures_1.59.0
[31] S4Arrays_1.7.1           tools_4.5.0
[33] parallel_4.5.0           BiocParallel_1.41.0
[35] memoise_2.0.1            GenomeInfoDbData_1.2.13
[37] curl_5.2.3               png_0.1-8
[39] vctrs_0.6.5              R6_2.5.1
[41] BiocIO_1.17.0            rtracklayer_1.67.0
[43] KEGGREST_1.47.0          zlibbioc_1.53.0
[45] bit_4.5.0                highr_0.11
[47] GenomicAlignments_1.43.0 xfun_0.49
[49] rjson_0.2.23             htmltools_0.5.8.1
[51] rmarkdown_2.29           compiler_4.5.0
```

```
[53] RCurl_1.98-1.16
```