

# COSMIC 67

*Julian Gehring, EMBL Heidelberg*

October 31, 2017

## 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
```

```
Loading required package: parallel
```

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

## COSMIC 67

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

*clusterApply, clusterApplyLB, clusterCall,  
clusterEvalQ, clusterExport, clusterMap, parApply,  
parCapply, parLapply, parLapplyLB, parRapply,  
parSapply, parSapplyLB*

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,  
append, as.data.frame, cbind, colMeans, colSums,  
colnames, do.call, duplicated, eval, evalq, get,  
grep, grepl, intersect, is.unsorted, lapply, lengths,  
mapply, match, mget, order, paste, pmax, pmax.int,  
pmin, pmin.int, rank, rbind, rowMeans, rowSums,  
rownames, sapply, setdiff, sort, table, tapply,  
union, unique, unsplit, which, which.max, which.min*

Loading required package: *GenomeInfoDb*

Loading required package: *S4Vectors*

Loading required package: *stats4*

Attaching package: '*S4Vectors*'

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

*expand.grid*

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")'.*

Loading required package: *DelayedArray*

Loading required package: *matrixStats*

Attaching package: '*matrixStats*'

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

*anyMissing, rowMedians*

## COSMIC 67

Attaching package: 'DelayedArray'

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

colMaxs, colMins, colRanges, rowMaxs, rowMins,  
rowRanges

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

apply

Loading required package: Rsamtools

Loading required package: Biostrings

Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:DelayedArray':

type

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

## COSMIC 67

```
geno(vcf):
  SimpleList 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 3.4.2 (2017-09-28)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.3 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.6-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.6-bioc/R/lib/libRlapack.so
```

```
locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8      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
```

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

## COSMIC 67

other attached packages:

```
[1] VariantAnnotation_1.24.0   Rsamtools_1.30.0
[3] Biostrings_2.46.0         XVector_0.18.0
[5] SummarizedExperiment_1.8.0 DelayedArray_0.4.0
[7] matrixStats_0.52.2       Biobase_2.38.0
[9] GenomicRanges_1.30.0     GenomeInfoDb_1.14.0
[11] IRanges_2.12.0           S4Vectors_0.16.0
[13] BiocGenerics_0.24.0      knitr_1.17
```

loaded via a namespace (and not attached):

```
[1] Rcpp_0.12.13             highr_0.6
[3] compiler_3.4.2          prettyunits_1.0.2
[5] progress_1.1.2          GenomicFeatures_1.30.0
[7] bitops_1.0-6            tools_3.4.2
[9] zlibbioc_1.24.0         biomaRt_2.34.0
[11] digest_0.6.12           bit_1.1-12
[13] BSgenome_1.46.0         memoise_1.1.0
[15] evaluate_0.10.1         RSQLite_2.0
[17] tibble_1.3.4            lattice_0.20-35
[19] rlang_0.1.2            Matrix_1.2-11
[21] DBI_0.7                 yaml_2.1.14
[23] GenomeInfoDbData_0.99.1 rtracklayer_1.38.0
[25] stringr_1.2.0           rprojroot_1.2
[27] bit64_0.9-7            grid_3.4.2
[29] R6_2.2.2                AnnotationDbi_1.40.0
[31] XML_3.98-1.9           RMySQL_0.10.13
[33] BiocParallel_1.12.0    rmarkdown_1.6
[35] blob_1.1.0             magrittr_1.5
[37] GenomicAlignments_1.14.0 backports_1.1.1
[39] htmltools_0.3.6        assertthat_0.2.0
[41] BiocStyle_2.6.0        stringi_1.1.5
[43] RCurl_1.95-4.8
```