

# Package ‘gwascat’

October 8, 2014

**Title** representing and modeling data in the NHGRI GWAS catalog

**Version** 1.8.0

**Author** VJ Carey <stvjc@channing.harvard.edu>

**Description** representing and modeling data in the NHGRI GWAS catalog

**Enhances** SNPlocs.Hsapiens.dbSNP.20120608, pd.genomewidesnp.6

**Depends** R (>= 3.0.0)

**Imports**

methods, Biostrings, IRanges, GenomicRanges, BiocGenerics, snpStats, Rsamtools, rtracklayer

**Suggests** DO.db, Gviz, ggbio, graph

**Maintainer** VJ Carey <stvjc@channing.harvard.edu>

**License** Artistic-2.0

**LazyData** yes

**biocViews** Genetics

## R topics documented:

gwascat-package	2
bindcadd_snv	3
gwastagger	4
gwaswloc-class	5
gwcex2gviz	7
gwdf_2012_02_02	7
locon6	9
makeCurrentGwascat	10
obo2graphNEL	11
riskyAlleleCount	12
topTraits	13
traitsManh	14

<b>Index</b>	<b>15</b>
--------------	-----------

---

gwascapackage      *representing and modeling data in the NHGRI GWAS catalog*

---

## Description

representing and modeling data in the NHGRI GWAS catalog, using GRanges and allied infrastructure

## Details

Package: gwascap  
Version: 1.7.3  
Suggests:  
Depends: R (>= 3.0.0), methods, IRanges, GenomicRanges  
Imports:  
License: Artistic-2.0  
LazyLoad: yes

### Index:

gwaswloc-class      Class "gwaswloc"

Upon attachment, a [GRanges-class](#) structure call gwrngs is formed which can be interrogated by position or through use of element metadata to learn about catalogued GWAS associations.

The data objects

'g17SM' 'gg17N' 'gw6.rs\_17' 'low17' 'rules\_6.0\_1kg\_17' 'gwrngs'

are described in vignettes.

The DataFrame function is imported from IRanges.

The [SnpMatrix-class](#) is used to represent data related to rule-based imputation, using the [impute.snps](#) function.

## Author(s)

VJ Carey <stvjc@channing.harvard.edu>

Maintainer: VJ Carey <stvjc@channing.harvard.edu>

## References

<http://www.genome.gov/gwastudies/>.

Partial support from the Computational Biology Group at Genentech, Inc.

**Examples**

```
## Not run:  
gwrngs  
  
## End(Not run)
```

---

`bindcadd_snv`*bind CADD scores of Kircher et al. 2014 to a GRanges instance*

---

**Description**

bind CADD scores of Kircher et al. 2014 to a GRanges instance; by default will use HTTP access at UW

**Usage**

```
bindcadd_snv(gr, fn = "http://krishna.gs.washington.edu/download/CADD/v1.0/1000G.tsv.gz")
```

**Arguments**

<code>gr</code>	query ranges to which CADD scores should be bound
<code>fn</code>	path to Tabix-indexed bgzipped TSV of CADD as distributed at krishna.gs.washington.edu on 1 April 2014

**Details**

joins CADD fields at addresses that match query; the CADD fields for query ranges that are not matched are set to NA

**Value**

GRanges instance with additional fields as obtained in the CADD resource

**Note**

This software developed in part with support from Genentech, Inc.

**Author(s)**

VJ Carey <stvjc@channing.harvard.edu>

**References**

M Kircher, DM Witten, P Jain, BJ O’Roak, GM Cooper, J Shendure, A general framework for estimating the relative pathogenicity of human genetic variants, Nature Genetics Feb 2014, PMID 24487276

**Examples**

```
## Not run:
# requires internet access
g2 = as(gwrngs, "GRanges")
bindcadd_snv( g2[which(seqnames(g2)=="chr2")][1:20] )

## End(Not run)
```

---

gwastagger

*data on 1000 genomes SNPs that 'tag' GWAS catalog entries*


---

**Description**

data on 1000 genomes SNPs that 'tag' GWAS catalog entries

**Usage**

```
data(gwastagger)
```

**Format**

The format is:

```
Formal class 'GRanges' [package "GenomicRanges"] with 6 slots
..@ seqnames :Formal class 'Rle' [package "IRanges"] with 4 slots
.. ..@ values : Factor w/ 24 levels "chr1","chr2",...: 1 2 3 4 5 6 7 8 9 10 ...
.. ..@ lengths : int [1:22] 24042 23740 21522 14258 14972 34101 12330 11400 8680 15429 ...
.. ..@ elementMetadata: NULL
.. ..@ metadata : list()
..@ ranges :Formal class 'IRanges' [package "IRanges"] with 6 slots
.. ..@ start : int [1:297579] 986111 988364 992250 992402 995669 999686 1005579 1007450
1011209 1011446 ...
.. ..@ width : int [1:297579] 1 1 1 1 1 1 1 1 1 1 ...
.. ..@ NAMES : NULL
.. ..@ elementType : chr "integer"
.. ..@ elementMetadata: NULL
.. ..@ metadata : list()
..@ strand :Formal class 'Rle' [package "IRanges"] with 4 slots
.. ..@ values : Factor w/ 3 levels "+","-","*": 3
.. ..@ lengths : int 297579
.. ..@ elementMetadata: NULL
.. ..@ metadata : list()
..@ elementMetadata:Formal class 'DataFrame' [package "IRanges"] with 6 slots
.. ..@ rownames : NULL
.. ..@ nrows : int 297579
.. ..@ listData :List of 3
.. .. ..$ tagid : chr [1:297579] "rs28479311" "rs3813193" "chr1:992250" "rs60442576" ...
.. .. ..$ R2 : num [1:297579] 0.938 0.994 0.969 1 1 ...
```

```

.. .. ..$ baseid: chr [1:297579] "rs3934834" "rs3934834" "rs3934834" "rs3934834" ...
.. .. ..@ elementType : chr "ANY"
.. .. ..@ elementMetadata: NULL
.. .. ..@ metadata : list()
..@ seqinfo :Formal class 'Seqinfo' [package "GenomicRanges"] with 4 slots
.. .. ..@ seqnames : chr [1:24] "chr1" "chr2" "chr3" "chr4" ...
.. .. ..@ seqlengths : int [1:24] 249250621 243199373 198022430 191154276 180915260 171115067
159138663 146364022 141213431 135534747 ...
.. .. ..@ is_circular: logi [1:24] FALSE FALSE FALSE FALSE FALSE FALSE ...
.. .. ..@ genome : chr [1:24] "hg19" "hg19" "hg19" "hg19" ...
..@ metadata : list()

```

## Details

This GRanges instance includes locations for 297000 1000 genomes SNP that have been identified as exhibiting LD with NHGRI GWAS SNP as of September 2013. The tagid field tells the name of the tagging SNP, the baseid field is the SNP identifier for the GWAS catalog entry, the R2 field tells the value of R-squared relating the distributions of the tagging SNP and the GWAS entry. Only tagging SNP with R-squared 0.8 or greater are included. A self-contained R-based procedure should emerge in 2014.

## Source

NHGRI GWAS catalog; plink is used with the 1000 genomes VCF in a perl routine by Michael McGeachie, Harvard Medical School;

## Examples

```

data(gwastagger)
gwastagger[1:5]
mean(gwrngs$SNPs %in% gwastagger$baseid)
# ideally, all GWAS SNP would be in our tagging ranges as baseid
query <- setdiff(gwrngs$SNPs, gwastagger$baseid)
# relatively recent catalog additions
sort(table(gwrngs[which(gwrngs$SNPs %in% query)]$Date.Added.to.Catalog), decreasing=TRUE)[1:10]

```

---

gwaswloc-class	<i>Class "gwaswloc"</i>
----------------	-------------------------

---

## Description

A container for GRanges instances representing information in the NHGRI GWAS catalog.

## Objects from the Class

Objects can be created by calls of the form `new("gwaswloc", ...)`. Any GRanges instance can be supplied.

**Slots**

extractDate: character set manually in .onAttach code to indicate date of retrieval of base table  
 seqnames: Object of class "Rle" typically representing chromosome numbers of loci associated with specific traits  
 ranges: Object of class "IRanges" genomic coordinates for locus  
 strand: Object of class "Rle" identifier of chromosome strand  
 elementMetadata: Object of class "DataFrame" general [DataFrame-class](#) instance providing attributes for the locus-trait association  
 seqinfo: Object of class "Seqinfo"  
 metadata: Object of class "list"

**Extends**

Class "[GRanges](#)", directly. Class "[GenomicRanges](#)", by class "[GRanges](#)", distance 2. Class "[Vector](#)", by class "[GRanges](#)", distance 3. Class "[GenomicRangesORmissing](#)", by class "[GRanges](#)", distance 3. Class "[GenomicRangesORGRangesList](#)", by class "[GRanges](#)", distance 3. Class "[Annotated](#)", by class "[GRanges](#)", distance 4.

**Methods**

[ signature(x = "gwaswloc"): a character argument to the bracket will be assumed to be a dbSNP identifier for a SNP locus, and records corresponding to this SNP are extracted; numeric indexes are supported as for [GRanges-class](#) instances.  
**getRsids** signature(x = "gwaswloc"): extract all dbSNP identifiers as a character vector  
**getTraits** signature(x = "gwaswloc"): extract all traits (NHGRI term 'Disease/Trait') as a character vector  
**subsetByChromosome** signature(x = "gwaswloc"): select records by chromosome, a vector of chromosomes may be supplied  
**subsetByTraits** signature(x = "gwaswloc"): select all records corresponding to a given vector of traits

**Note**

In gwascat package, the globally accessible gwaswloc instance gwrngs is created upon attachment.

**Author(s)**

VJ Carey <stvjc@channing.harvard.edu>

**References**

<http://www.genome.gov/gwastudies/>

**Examples**

```
showClass("gwaswloc")
```

---

gwce2gviz	<i>Prepare salient components of GWAS catalog for rendering with Gviz</i>
-----------	---

---

**Description**

Prepare salient components of GWAS catalog for rendering with Gviz

**Usage**

```
gwce2gviz(basegr = gwrngs, contextGR = GRanges(seqnames =
  "chr17", IRanges(start = 37500000, width = 1e+06)),
  txrefpk = "TxDb.Hsapiens.UCSC.hg19.knownGene", genome
  = "hg19", genesympk = "org.Hs.eg.db", plot.it = TRUE,
  maxmlp = 25)
```

**Arguments**

basegr	gwaswloc instance containing information about GWAS in catalog
contextGR	A GRanges instance delimiting the visualization in genomic coordinates
txrefpk	a TxDb package, typically
genesympk	string naming annotationDbi .db package
genome	character tag like 'hg19'
plot.it	logical, if FALSE, just return list
maxmlp	maximum value of $-10 \log p$ – winsorization of all larger values is performed, modifying the contents of Pvalue_mlogp in the elementMetadata for the call

**Examples**

```
args(gwce2gviz)
#gwascat:::onAttach("", "gwascat")
gwce2gviz()
```

---

gwdf_2012_02_02	<i>internal data frame for NHGRI GWAS catalog</i>
-----------------	---

---

**Description**

convenience container for imported table from NHGRI GWAS catalog

**Usage**

```
data(gwdf_2012_09_22) # or more recent elements available
```

**Format**

A data frame with 9000+ observations on the following 34 variables.

Date Added to Catalog a character vector  
PUBMEDID a character vector  
First Author a character vector  
Date a character vector  
Journal a character vector  
Link a character vector  
Study a character vector  
Disease/Trait a character vector  
Initial Sample Size a character vector  
Replication Sample Size a character vector  
Region a character vector  
Chr\_id a character vector  
Chr\_pos a character vector  
Reported Gene(s) a character vector  
Mapped\_gene a character vector  
Upstream\_gene\_id a character vector  
Downstream\_gene\_id a character vector  
Snp\_gene\_ids a character vector  
Upstream\_gene\_distance a character vector  
Downstream\_gene\_distance a character vector  
Strongest SNP-Risk Allele a character vector  
SNPs a character vector  
Merged a character vector  
Snp\_id\_current a character vector  
Context a character vector  
Intergenic a character vector  
Risk Allele Frequency a character vector  
p-Value a character vector  
Pvalue\_mlog a character vector  
p-Value (text) a character vector  
OR or beta a character vector  
95% CI (text) a character vector  
Platform. . a character vector  
CNV a character vector



**Note**

The `.onAttach` function specifies which data frame is transformed to GRanges.

**Source**

<http://www.genome.gov/gwastudies>

**Examples**

```
## Not run:  
data(gwdf_2012_03_22)  
  
## End(Not run)
```

---

locon6	<i>location information for 10000 SNPs probed on Affy GW 6.0</i>
--------	--

---

**Description**

location information for 10000 SNPs probed on Affy GW 6.0

**Usage**

```
data(locon6)
```

**Format**

A data frame with 10000 observations on the following 3 variables.

`db SNP_rs_id` a character vector

`chrom` a character vector

`physical_pos` a numeric vector

**Details**

extracted from `pd.genomewidesnp.6 v 1.4.0`; for demonstration purposes

**Examples**

```
data(locon6)  
str(locon6)
```

---

makeCurrentGwascat	<i>read NHGRI GWAS catalog table and construct associated GRanges instance</i>
--------------------	--

---

**Description**

read NHGRI table and construct associated GRanges instance

**Usage**

```
makeCurrentGwascat(table.url = "http://www.genome.gov/admin/gwascatalog.txt", fixNonASCII = TRUE)
```

**Arguments**

table.url	string identifying the .txt file curated at NHGRI
fixNonASCII	logical, if TRUE, non-ASCII characters as identified by iconv will be replaced by asterisk

**Details**

records for which clear genomic position cannot be determined are dropped from the ranges instance  
an effort is made to use reasonable data types for GRanges metadata, so some qualifying characters such as (EA) in Risk allele frequency field will simply be omitted during coercion of contents of that field to numeric.

**Value**

a GRanges instance

**Author(s)**

VJ Carey

**Examples**

```
## Not run:  
# if you have good internet access  
newcatr = makeCurrentGwascat()  
  
## End(Not run)
```

---

obo2graphNEL	<i>convert a typical OBO text file to a graphNEL instance (using Term elements)</i>
--------------	---

---

**Description**

convert a typical OBO text file to a graphNEL instance (using Term elements)

**Usage**

```
obo2graphNEL(obo, kill = "\\[Typedef\\]")
```

**Arguments**

obo	string naming a file in OBO format
kill	entity types to be excluded from processing – probably this should be in a 'keep' form, but for now this works.

**Details**

Very rudimentary list and grep operations are used to retain sufficient information to map the DAG to a graphNEL, using formal term identifiers as node names and 'is-a' relationships as edges, and term names and other metadata are assigned to nodeData components.

**Value**

a graphNEL instance

**Note**

The OBO for Human Disease ontology is serialized as text with this package.

**Author(s)**

VJ Carey <stvjc@channing.harvard.edu>

**References**

For use with human disease ontology, [http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease\\_ontology](http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology)

---

riskyAlleleCount	<i>given a matrix of subjects x SNP calls, count number of risky alleles</i>
------------------	--

---

### Description

given a matrix of subjects x SNP calls, count number of risky alleles for various conditions, relative to NHGRI GWAS catalog

### Usage

```
riskyAlleleCount(callmat, matIsAB = TRUE, chr,
  gwml = gwrngs, snpap = "SNPlocs.Hsapiens.dbSNP.20120608",
  gencode = c("A/A", "A/B", "B/B"))
```

### Arguments

callmat	matrix with subjects as rows, SNPs as columns; entries can be generic A/A, A/B, B/B, or specific nucleotide calls
matIsAB	logical, FALSE if nucleotide codes are present, TRUE if generic call codes are present; in the latter case, gwascats:::ABmat2nuc will be run
chr	code for chromosome, should work with the SNP annotation getSNPlocs function, so likely "ch[nn]"
gwml	an instance of <a href="#">gwaswloc</a>
snpap	name of a Bioconductor SNPlocs.Hsapiens.dbSNP.* package
gencode	codes used for generic SNP call

### Value

matrix with rows corresponding to subjects , columns corresponding to SNP

### Examples

```
#if (!exists("gwrngs")) gwascats:::onAttach("a", "b")
data(gg17N) # translated from GGdata chr 17 calls using ABmat2nuc
h17 = riskyAlleleCount(gg17N, matIsAB=FALSE, chr="ch17")
h17[1:5,1:5]
table(as.numeric(h17))
```

---

topTraits                      *operations on GWAS catalog*

---

**Description**

operations on GWAS catalog

**Usage**

```
topTraits (gwwl, n=10, tag="Disease.Trait")
locs4trait(gwwl, trait, tag="Disease.Trait")
chklocs(chrtag="20", gwwl=gwrngs)
```

**Arguments**

gwwl	instance of <a href="#">gwaswloc</a>
n	numeric, number of traits to report
tag	character, name of field to be used for trait enumeration
trait	character, trait to use for filtering
chrtag	character, chromosome identifier

**Value**

topTraits returns a character vector of most frequently occurring traits in the database

locs4trait returns a [gwaswloc](#) object with records defining associations to the specified trait

chklocs returns a logical that is TRUE when the asserted locations of SNP in the GWAS catalog agree with the locations given in the dbSNP package `SNPlocs.Hsapiens.dbSNP.20110815`

**Author(s)**

VJ Carey <stvjc@channing.harvard.edu>

**Examples**

```
#if (!exists("gwrngs")) gwascat:::onAttach("a", "b")
topTraits(gwrngs)
```

---

traitsManh	<i>use ggbio facilities to display GWAS results for selected traits in genomic coordinates</i>
------------	--

---

**Description**

use ggbio facilities to display GWAS results for selected traits in genomic coordinates

**Usage**

```
traitsManh(gwr, selr = GRanges(seqnames = "chr17", IRanges(3e+07, 5e+07)), traits = c("Asthma", "Parkin
```

**Arguments**

<code>gwr</code>	GRanges instance as managed by the gwaswloc container design, with Disease.Trait and Pvalue\_mlog among elementMetadata columns
<code>selr</code>	A GRanges instance to restrict the gwr for visualization. Not tested for noncontiguous regions.
<code>traits</code>	Character vector of traits to be exhibited; GWAS results with traits not among these will be labeled "other".
<code>truncmlp</code>	Maximum value of $-\log_{10} p$ to be displayed; in the raw data this ranges to the hundreds and can cause bad compression.
<code>...</code>	not currently used

**Details**

uses a ggbio autoplot

**Value**

autoplot value

**Author(s)**

VJ Carey <stvjc@channing.harvard.edu>

**Examples**

```
# do a p-value truncation if you want to reduce compression
#gwascat:::onAttach("A", "gwascat")
traitsManh(gwrngs)
```

# Index

- \*Topic **classes**
  - gwaswloc-class, 5
- \*Topic **datasets**
  - gwastagger, 4
  - gwdf\_2012\_02\_02, 7
  - locon6, 9
- \*Topic **graphics**
  - gwce2gviz, 7
  - traitsManh, 14
- \*Topic **models**
  - bindcadd\_snv, 3
  - makeCurrentGwascat, 10
  - obo2graphNEL, 11
  - riskyAlleleCount, 12
  - topTraits, 13
  - traitsManh, 14
- \*Topic **package**
  - gwascat-package, 2
- [, gwaswloc, ANY, ANY, ANY-method
  - (gwaswloc-class), 5
- [, gwaswloc, ANY-method (gwaswloc-class), 5
- [, gwaswloc-method (gwaswloc-class), 5
- Annotated, 6
- bindcadd\_snv, 3
- chklocs (topTraits), 13
- DataFrame (gwascat-package), 2
- g17SM (gwascat-package), 2
- GenomicRanges, 6
- GenomicRangesORGRangesList, 6
- GenomicRangesORmissing, 6
- getRsids (gwaswloc-class), 5
- getRsids, gwaswloc-method
  - (gwaswloc-class), 5
- getTraits (gwaswloc-class), 5
- getTraits, gwaswloc-method
  - (gwaswloc-class), 5
- gg17N (gwascat-package), 2
- GRanges, 6
- gw6.rs\_17 (gwascat-package), 2
- gwascat (gwascat-package), 2
- gwascat-package, 2
- gwastagger, 4
- gwaswloc, 12, 13
- gwaswloc-class, 5
- gwce2gviz, 7
- gwdf\_2012\_02\_02, 7
- gwdf\_2012\_03\_20 (gwdf\_2012\_02\_02), 7
- gwdf\_2012\_09\_22 (gwdf\_2012\_02\_02), 7
- gwrngs (gwascat-package), 2
- impute.snps, 2
- impute.snps (gwascat-package), 2
- locon6, 9
- locs4trait (topTraits), 13
- low17 (gwascat-package), 2
- makeCurrentGwascat, 10
- obo2graphNEL, 11
- riskyAlleleCount, 12
- rules\_6.0\_1kg\_17 (gwascat-package), 2
- Snpmatrix-class (gwascat-package), 2
- subsetByChromosome (gwaswloc-class), 5
- subsetByChromosome, gwaswloc-method
  - (gwaswloc-class), 5
- subsetByTraits (gwaswloc-class), 5
- subsetByTraits, gwaswloc-method
  - (gwaswloc-class), 5
- topTraits, 13
- traitsManh, 14
- Vector, 6