Package ‘gwascat’

January 21, 2017

Title representing and modeling data in the EMBL-EBI GWAS catalog

Version 2.6.0

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

Description Represent and model data in the EMBL-EBI GWAS catalog.

Enhances SNPloc.Hsapiens.dbSNP144.GRCh37

Depends R (>= 3.0.0), Homo.sapiens

Imports methods, BiocGenerics, S4Vectors (>= 0.9.25), IRanges, GenomeInfoDb, GenomicRanges, snpStats, Biostrings, Rsamtools, rtracklayer, gQTL.stats, Gviz, VariantAnnotation, AnnotationHub, AnnotationDbi, GenomicFeatures, graph, ggbio, ggplot2, SummarizedExperiment

Suggests DO.db, DT, utils, knitr, RBGL, RUnit, GGtools

VignetteBuilder utils, knitr

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

License Artistic-2.0

LazyData no

biocViews Genetics

NeedsCompilation no

R topics documented:

- gwascat-package .......................................................... 2
- bindcadd_snv ............................................................ 3
- gwastagger ............................................................... 4
- gwaswloc-class .......................................................... 5
- gwcex2gviz ............................................................... 7
- gwdf_2012_02_02 ...................................................... 7
- ldtagr .................................................................. 9
- locon6 ................................................................. 10
- makeCurrentGwascat .................................................... 11
- obo2graphNEL ........................................................... 12
- riskyAlleleCount ......................................................... 13
- topTraits ................................................................. 14
- traitsManh ............................................................... 15

Index 16
representing and modeling data in the NHGRI GWAS catalog, using GRanges and allied infrastructure

Details

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

Index:

- **gwaswloc-class**: Class "gwaswloc"

The GWAS catalog management has migrated to EMBL/EBI. Use data(ebicat38) for an image dated 3 August 2015. Use makeCurrentGwascat() to get a more recent image. Use data(ebicat37) for a GRCh37 (or hg19) liftOver result. Use data(ebicat37UCSC) for an image with hg19 as genome tag and UCSC seqnames.

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.

si.hs.38 is a **Seqinfo-class** instance for hg38.

**nodeData** (and nodes, ugraph, subGraph, adj) are exported for use in the vignettes.

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.
bindcadd_snv

Examples

```r
## Not run:
data(ebicat38)
ebicat38
## End(Not run)
```

### Description

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

#### Usage

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

#### Arguments

- `gr`: query ranges to which CADD scores should be bound
- `fn`: 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

Examples

```r
## Not run:
# requires internet access
data(ebicat37)
g2 = as(ebicat37, "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 1 1 1 1 1...
  .. .. ..$ baseid: chr [1:297579] "rs3934834" "rs3934834" "rs3934834" "rs3934834" "rs3934834"...
  .. ..@ elementType : chr "ANY"
  .. ..@ elementMetadata: NULL
```
gwaswloc-class

.. .. ..@ 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 GW AS 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]
data(ebicat37)
mean(ebicat37$SNPS %in% gwastagger$baseid)
# ideally, all GWAS SNP would be in our tagging ranges as baseid
query <- setdiff(ebicat37$SNPS, gwastagger$baseid)
# relatively recent catalog additions
sort(table(ebicat37[which(ebicat37$SNPS %in% query)]$DATE.ADDED.TO.CATALOG), decreasing=TRUE)[1:10]

gwaswloc-class

Class "gwaswloc"

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


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 1.9.6 and earlier, the globally accessible gwaswloc instance gwrngs was created upon attachment. This is no longer the case.

Author(s)

VJ Carey <stvjc@channing.harvard.edu>

References

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

Examples

showClass("gwaswloc")
**gwcex2gviz**

Prepare salient components of GWAS catalog for rendering with Gviz

**Description**

Prepare salient components of GWAS catalog for rendering with Gviz

**Usage**

```r
gwcex2gviz(basegr, contextGR = GRanges(seqnames = "chr17", IRanges(start = 37500000, width = 1e+06)), txrefpk = "TxDb.Hsapiens.UCSC.hg19.knownGene", genome = "hg19", genesympk = "Homo.sapiens", 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**

```r
args(gwcex2gviz)
data("gwdf_2014_09_08")
```

**gwdf_2012_02_02**

Internal data frame for NHGRI GWAS catalog

**Description**

Convenience container for imported table from NHGRI GWAS catalog

**Usage**

```r
data("gwdf_2014_09_08")
```
Format

A data frame with 17832 observations on the following 34 variables.

'Date Added to Catalog' a character vector
PUBMEDID a numeric 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
'SPlatform [SNPs passing QC]' a character vector
CNV a character vector

Note

In versions prior to 1.9.6, the `.onAttach` function specifies which data frame is transformed to GRanges. This is now managed manually.
ldtagr

Source
http://www.genome.gov/gwastudies

Examples

```r
## Not run:
data(gwdf_2014_09_08)
# try gwascat:::gwdf2GRanges on this data.frame
## End(Not run)
```

ldtagr expand a list of variants by including those in a VCF with LD exceeding some threshold

Description

expand a list of variants by including those in a VCF with LD exceeding some threshold

Usage

```r
ldtagr(snprng, tf, samples, genome = "hg19", lbmaf = 0.05, lbR2 = 0.8, radius = 1e+05)
```

Arguments

- **snprng**: a named GRanges for a single SNP. The name must correspond to the name that will be assigned by `genotypeToSnpMatrix` to the corresponding column of a SnpMatrix.
- **tf**: TabixFile instance pointing to a bgzipped tabix-indexed VCF file
- **samples**: a vector of sample identifiers, if excluded, all samples used
- **genome**: tag like 'hg19'
- **lbmaf**: lower bound on variant MAF to allow consideration
- **lbR2**: lower bound on R squared for regarding SNP to be incorporated
- **radius**: radius of search in bp around the input range

Details

uses snpStats ld()

Value

a GRanges with names corresponding to 'new' variants and mcols fields 'paramRangeID' (base variant input) and 'R2'

Note

slow but safe approach. probably a matrix method could be substituted using the nice sparse approach already in snpStats
Author(s)

VJ Carey

Examples

```r
require(GenomicRanges)
cand = GRanges("1", IRanges(113038694, width=1))
names(cand) = "rs883593"
require(VariantAnnotation)
expath = dir(system.file("vcf", package="GGtools"), patt=".*exon.*gz\$", full=TRUE)
tf = TabixFile(expath)
ldtagr( cand, tf, lbR2 = .8)
# should do with 1000 genomes in S3 bucket and gwascat
```

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.

- `dbsnp_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

```r
data(locon6)
str(locon6)
```
**makeCurrentGwascat**

*read NHGRI GWAS catalog table and construct associated GRanges instance*

**Description**

read NHGRI table and construct associated GRanges instance

**Usage**

```r
makeCurrentGwascat(table.url =
        "http://www.ebi.ac.uk/gwas/api/search/downloads/alternative",
        fixNonASCII = TRUE, genome="GRCh38",
        withOnt = TRUE)
```

**Arguments**

- `table.url` string identifying the .txt file curated at EBI/EMBL
- `fixNonASCII` logical, if TRUE, non-ASCII characters as identified by iconv will be replaced by asterisk
- `genome` character string: 'GRCh38' is default and yields current image as provided by EMBL/EBI; 'GRCh37' yields a realtime liftOver to hg19 coordinates, via AnnotationHub storage of the chain files. Any other value yields an error.
- `withOnt` logical indicating whether 'alternative' (ontology-present, includes repetition of loci with one:many ontological mapping) or 'full' (ontology-absent, one record per locus report) version of distributed table

**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**

```r
## Not run:
# if you have good internet access
ewcatr = makeCurrentGwascat()

## End(Not run)
```
obo2graphNEL

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

Description

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

Usage

obo2graphNEL(obo, kill = "\[Typedef\]", killTrailSp=TRUE)
node2uri(nn)
uri2node(us)

Arguments

obo string naming a file in OBO format
nn node name for converted graph, generally of form EFO:nnnnnn
us URI string from GWAS catalog annotation MAPPED\_TRAIT\_URI
kill entity types to be excluded from processing – probably this should be in a 'keep' form, but for now this works.
killTrailSp In the textual version of EFO ca. Aug 2015, there is a trailing blank in the tag field defining EFO:0000001, which is not present in references to this term. Set this to TRUE to eliminate this, or graphNEL construction will fail to validate.

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
riskyAlleleCount

Examples

data(efo.obo.g)
hn = nodes(efo.obo.g)[1:5]
hn
nodeData(efo.obo.g, hn[5])

riskyAlleleCount
given a matrix of subjects x SNP calls, count number of risky alleles

Description

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

Usage

riskyAlleleCount(callmat, matIsAB = TRUE, chr,
gwwl, snpap = "SNPlocs.Hsapiens.dbSNP144.GRCh37",
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, gwascat:::ABmat2nuc will be run
chr code for chromosome, should work with the SNP annotation getSNPlocs function, so likely "ch[nn]"
gwwl an instance of gwaswloc
.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

data(gg17N) # translated from GGdata chr 17 calls using ABmat2nuc
data(ebicat37)
library(GenomeInfoDb)
seqlevelsStyle(ebicat37) = "UCSC"
h17 = riskyAlleleCount(gg17N, matIsAB=FALSE, chr="ch17", gwwl=ebicat37)
h17[1:5,1:5]
table(as.numeric(h17))
## Description

operations on GWAS catalog

## Usage

```r
topTraits (gwwl, n=10, tag="DISEASE/TRAIT")
locs4trait(gwwl, trait, tag="DISEASE/TRAIT")
chklocs(chrtag="20", gwwl)
```

## Arguments

- `gwwl`: instance of `gwaswloc`
- `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.dbSNP144.GRCh37

## Author(s)

VJ Carey <stvjc@channing.harvard.edu>

## Examples

```r
data(ebicat38)
topTraits(ebicat38)
```
traitsManh

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

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", "Parkinson's disease", "Height", "Crohn's disease"), truncmlp = 25, ...)

Arguments

gwr GRanges instance as managed by the gwaswloc container design, with Disease.Trait and Pvalue\_mlog among elementMetadata columns

selr A GRanges instance to restrict the gwr for visualization. Not tested for noncontiguous regions.

traits Character vector of traits to be exhibited; GWAS results with traits not among these will be labeled “other”.

truncmlp Maximum value of -log10 p to be displayed; in the raw data this ranges to the hundreds and can cause bad compression.

... not currently used

Details

uses a ggbio autoplot

Value

autoplot value

Note

An xlab is added, concatenating genome tag with seqnames tag.

Author(s)

VJ Carey <stvjc@channing.harvard.edu>

Examples

# do a p-value truncation if you want to reduce compression
data(ebicat38)
library.GenomeInfoDb()
seqlevelsStyle(ebicat38) = "UCSC"
traitsManh(ebicat38)
Index

*Topic classes
  gwaswloc-class, 5

*Topic datasets
  gwastagger, 4
  gwdf_2012_02_02, 7
  locon6, 10

*Topic graphics
  gwcex2gviz, 7
  traitsManh, 15

*Topic models
  bindcadd_snv, 3
  ldtagr, 9
  makeCurrentGwascat, 11
  obo2graphNEL, 12
  riskyAlleleCount, 13
  topTraits, 14
  traitsManh, 15

*Topic package
  gwascat-package, 2
  [,gwaswloc,ANY,ANY,ANY-method (gwaswloc-class), 5
  [,gwaswloc,ANY-method (gwaswloc-class), 5
  [,gwaswloc-method (gwaswloc-class), 5
  adj (gwascat-package), 2
  Annotated, 6
  bindcadd_snv, 3
  chklocs (topTraits), 14
  DataFrame (gwascat-package), 2
  ebicat37 (gwascat-package), 2
  ebicat38 (gwascat-package), 2
  efo.obo.g (obo2graphNEL), 12
  g17SM (gwascat-package), 2
  GenomicRanges, 6
  GenomicRangesORGRangesList, 6
  GenomicRangesORmissing, 6
  genotypeToSnpMatrix, 9
  getRsids (gwaswloc-class), 5
  getRsids.gwaswloc-method (gwaswloc-class), 5
  getTraits (gwaswloc-class), 5
  getTraits (gwaswloc-class), 5
  getTraits (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, 13, 14
  gwaswloc-class, 5
  gwcex2gviz, 7
  gwdf_2012_02_02, 7
  gwdf_2014_09_08 (gwdf_2012_02_02), 7
  gwrngs19 (gwascat-package), 2
  gwrngs38 (gwascat-package), 2
  impute.snps, 2
  impute.snps (gwascat-package), 2
  ldtagr, 9
  locon6, 10
  locs4trait (topTraits), 14
  low17 (gwascat-package), 2
  makeCurrentGwascat, 11
  node2uri (obo2graphNEL), 12
  nodeData, 2
  nodeData (gwascat-package), 2
  nodes (gwascat-package), 2
  obo2graphNEL, 12
  riskyAlleleCount, 13
  rules_6.0_1kg_17 (gwascat-package), 2
  show, gwaswloc-method (gwaswloc-class), 5
  si.hs.38 (gwascat-package), 2
  SnpMatrix-class (gwascat-package), 2
  subGraph (gwascat-package), 2
  subsetByChromosome (gwaswloc-class), 5

16
INDEX

subsetByChromosome, gwaswloc-method (gwaswloc-class), 5
subsetByTraits (gwaswloc-class), 5
subsetByTraits, gwaswloc-method (gwaswloc-class), 5

topTraits, 14
traitsManh, 15

ugraph (gwascat-package), 2
uri2node (obo2graphNEL), 12

Vector, 6