Package 'gap.datasets'

May 9, 2022

Version 0.0.5

Date 2022-5-6

Title Datasets for 'gap'

Description Datasets associated with the 'gap' package. Currently, it includes an example data for regional association plot (CDKN), an example data for a genomewide association meta-analysis (OPG), data in studies of Parkinson's diease (PD), ALHD2 markers and alcoholism (aldh2), APOE/APOC1 markers and Schizophrenia (apoeapoc), cystic fibrosis (cf), a Olink/INF panel (inf1), Manhattan plots with (hr1420, mhtdata) and without (w4) gene annotations.

LazyData Yes

LazyLoad Yes

License GPL (>= 2)

URL https://jinghuazhao.github.io/R/

NeedsCompilation no

Depends R (>= 2.10)

Author Jing Hua Zhao [aut, cre], Swetlana Herbrandt [ctb]

Maintainer Jing Hua Zhao <jinghuazhao@hotmail.com>

Repository CRAN

Date/Publication 2022-05-09 08:10:02 UTC

R topics documented:

`aldh2` *ALDH2 markers and Alcoholism*

Description

This data set contains eight ALDH2 markers and Japanese alcohlic patients (y=1) and controls (y=0). There are genotypes for 8 loci, with a prefix name (e.g., "EXON12") and a suffix for each of two alleles (".a1" and ".a2").

The eight markers loci follows the following map (base pairs)

D12S2070	(> 450 000),
D12S839	(> 450 000),
D12S821	(∼ 400 000),
D12S1344	(83 853),
EXON12	(0),
EXON1	(37 335),
D12S2263	(38 927),
D12S1341	(> 450 000)

Usage

```
data(aldh2)
```

Format

A data frame

Source

Prof Ian Craig of Oxford and SGDP Centre, KCL

References

Koch HG, McClay J, Loh E-W, Higuchi S, Zhao J-H, Sham P, Ball D, et al (2000) Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. Hum. Mol. Genet. 9:2993-2999

---

apoeapoc *APOE/APOC1 markers and Alzheimer's*

Description

This data set contains APOE/APOC1 markers and Chinese Alzheimer's patients and controls. Variable id is subject id and y takes value 0 for controls and 2 for Alzheimer's.

The last six variables are age, sex and genotypes for APOE and APOC with suffixes for each of two alleles (".a1" and ".a2").

Usage

```
data(apoeapoc)
```

Format

A data frame

Source

Shi J, Zhang S, Ma C, Liu X, Li T, Tang M, Han H, Guo Y, Zhao JH, Zheng K, Kong X, Zhang K, Su Z, Zhao Z. Association between apolipoprotein CI HpaI polymorphism and sporadic Alzheimer's disease in Chinese. Acta Neurol Scan 2004, 109:140-145.

CDKN *An example data for regional association plot*

Description

These data are adapted from the DGI study on CDKN2A/CDKN2B region.

Usage

```
data(CDKN)
```

Format

There are three data objects in the dataset: CDKNgenes, the gene list from the Chromosome 9 according to UCSC browser (https://genome.ucsc.edu/); CDKNmap, the genetic map as from the HapMap website (https://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2006-10_rel21_phaseI+II/rates/); CDKNlocus, the results from the association analysis of the locus based on DGI data.

Source

The data were obtained from the Harvard-MIT Broad Institute (see https://www.broadinstitute.org/diabetes)

References

Diabetes Genetics Initiative of Broad Institute of Havard and MIT, Lund University and Novartis Institute for BioMedical Research. *Whole-genome association analysis identifies novel loci for type 2 diabetes and triglyceride levels* Science 2007;316(5829):1331-6

Examples

```
data(CDKN)
head(CDKNlocus)
```

---

cf *Cystic fibrosis data*

Description

This data set contains a case-control indicator and 23 SNPs.

The inter-marker distances (Morgan) are as follows

0.000090, 0.000158, 0.005000, 0.000100, 0.000200, 0.000150, 0.000250, 0.000200, 0.000050, 0.000350, 0.000300, 0.000250, 0.000350, 0.000350, 0.000800, 0.000100, 0.000200, 0.000150, 0.000550, 0.006000, 0.000700, 0.001000

Usage

```
data(cf)
```

Format

A data frame containing 186 rows and 24 columns

Note

This can be used as an example of converting PL-EM to matrix format,

```
cfdata <- vector("numeric")
cfname <- vector("character")
for (i in 2:dim(cf)[2])
{
    tmp <- plem2m(cf[,i])
    a1 <- tmp[[1]]
    a2 <- tmp[[2]]
    cfdata <- cbind(cfdata,a1,a2)
    a1name <- paste("loc",i-1,".a1",sep="")
    a2name <- paste("loc",i-1,".a2",sep="")
    cfname <- cbind(cfname,a1name,a2name)
}
cfdata <- as.data.frame(cfdata)
names(cfdata) <- cfname
```

Source

Liu JS, Sabatti C, Teng J, Keats BJB, Risch N (2001). Bayesian Analysis of Haplotypes for Linkage Disequilibrium Mapping. Genome Research 11:1716-1724

---

cnv *A CNV data*

Description

A CNV dataset.

Usage

```
data(cnv)
```

Format

A CNV data

Source

Zheng Ye

crohn *Crohn's disease data*

Description

The data set consist of 103 common (>5% minor allele frequency) SNPs genotyped in 129 trios from an European-derived population. These SNPs are in a 500-kb region on human chromosome 5q31 implicated as containing a genetic risk factor for Crohn disease.

The positions, names and haplotype blocks reported are as follows,

```
274044   IGR1118a_1 BLOCK 1
274541   IGR1119a_1 *
286593   IGR1143a_1 *
287261   IGR1144a_1 *
299755   IGR1169a_2 *
324341   IGR1218a_2 *
324379   IGR1219a_2 *
358048   IGR1286a_1 BLOCK 1
366811  TSC0101718
395079   IGR1373a_1 BLOCK 2
396353   IGR1371a_1 *
397334   IGR1369a_2 *
397381   IGR1369a_1 *
398352   IGR1367a_1 BLOCK 2
411823   IGR2008a_2
411873   IGR2008a_1 BLOCK 3
412456   IGR2010a_3 *
413233   IGR2011b_1 *
415579   IGR2016a_1 *
417617   IGR2020a_15 *
419845   IGR2025a_2 *
424283   IGR2033a_1 *
425376   IGR2036a_2 *
425549   IGR2036a_1 BLOCK 3
433467   IGR2052a_1 BLOCK 4
435282   IGR2055a_1 *
437682   IGR2060a_1 *
438883   IGR2063b_1 *
443565   IGR2072a_2 *
443750   IGR2073a_1 *
445337   IGR2076a_1 *
447791   IGR2081a_1 *
449895   IGR2085a_2 *
455246   IGR2096a_1 *
463136   IGR2111a_3 BLOCK 4
482171   IGR2150a_1 BLOCK 5
```

```
485828   IGR2157a_1 *
495082   IGR2175a_2 *
506266   IGR2198a_1 *
506890   IGR2199a_1 BLOCK 5
507208   IGR2200a_1 BLOCK 6
508338   IGR2202a_1 *
508858   IGR2203a_1 *
510951   IGR2207a_1 *
518478   IGR2222a_2 BLOCK 6
519387   IGR2224a_2 BLOCK 7
519962   IGR2225a_1 *
520521   IGR2226a_3 *
522600   IGR2230a_1 *
525243   IGR2236a_1 *
529556   IGR2244a_4 *
532363   IGR2250a_4 *
545062   IGR2276a_1 *
553189   IGR2292a_1 *
570978   IGR3005a_1 *
571022   IGR3005a_2 *
576586   IGR3016a_1 *
577141   IGR3018a_2 *
577838   IGR3019a_2 *
578122   IGR3020a_1 *
579217   IGR3022a_1 *
579529   IGR3023a_1 *
579818   IGR3023a_3 *
582651   IGR3029a_1 *
582948   IGR3029a_2 *
583131   IGR3030a_1 *
587836   IGR3039a_1 *
590425   IGR3044a_1 *
590585   IGR3045a_1 *
594115   IGR3051a_1 *
594812   IGR3053a_1 *
598805   IGR3061a_1 *
601294   IGR3066a_1 *
608759   IGR3081a_1 *
610447   IGR3084a_1 *
611177   IGR3086a_1 BLOCK 7
613488   IGR3090a_1
616241   IGR3096a_1 BLOCK 8
616763   IGR3097a_1 *
617299   IGR3098a_1 *
626881   IGR3117a_1 *
633786   IGR3131a_1 *
635072   IGR3134a_1 *
637441   IGR3138a_1 BLOCK 8
```

```
648564   IGR3161a_1
649061   IGR3162a_1 BLOCK 9
649903   IGR3163a_1 *
657234   IGR3178a_1 *
662077   IGR3188a_1 *
662819   IGR3189a_2 *
676688   IGRX100a_1 BLOCK 9
683387   IGR3230a_1 BLOCK 10
686249   IGR3236a_1 *
692320   IGR3248a_1 *
718291   IGR3300a_2 *
730313   IGR3324a_1 *
731025   IGR3326a_1 *
738461   IGR3340a_1 BLOCK 10
871978   GENS021ex1_2 BLOCK 11
877571   GENS020ex3_3 *
877671   GENS020ex3_2 *
877809   GENS020ex3_1 *
890710   GENS020ex1_1 BLOCK 11
```

However it has been changed after the paper was published.

An example use of the data is with the following paper, Kelly M. Burkett, Celia M. T. Greenwood, BradMcNeney, Jinko Graham. Gene genealogies for genetic association mapping, with application to Crohn's disease. Fron Genet 2013, 4(260) doi: 10.3389/fgene.2013.00260

Usage

```
data(crohn)
```

Format

A data frame containing 387 rows and 212 columns

Source

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001). High-resolution haplotype structure in the human genome Nature Genetics 29:229-232

`fa` *Friedreich Ataxia data*

Description

This data set contains a case-control indicator and twelve microsatellite markers. An extra unphased individual with the following genotype

```
 2  7  7  7  1  3  2  2  2  2  6  3
 3  8 10  8  3  9  3  4  2  2  7  5
```

has not been included.

The inter-marker distances (Morgan) are as follows,

0.03, 0.065, 0.00125, 0.00125, 0.00125, 0.00125, 0.00125, 0.00125, 0.00125, 0.00125, 0.045

Usage

```
data(fa)
```

Format

A data frame containing 127 rows and 13 columns

Source

Liu JS, Sabatti C, Teng J, Keats BJB, Risch N (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping Genome Research 11:1716-1724

---

fsnps *A case-control data involving four SNPs with missing genotype*

Description

This is a simulated data of four SNPs with their alleles coded in characters. The variable y contains phenotypes (1=case, 0=control).

Usage

```
data(fsnps)
```

Format

A data frame

Source

Dr Sebastien Lissarrague of Genset

hla *The HLA data*

Description

This data set contains HLA markers DRB, DQA, DQB and phenotypes of 271 Schizophrenia patients (y=1) and controls (y=0). Genotypes for 3 HLA loci have prefixes name (e.g., "DQB") and a suffix for each of two alleles (".a1" and ".a2").

Usage

```
data(hla)
```

Format

A data frame containing 271 rows and 8 columns

Source

Dr Padraig Wright of Pfizer

hr1420 *An example data for Manhattan plot with annotation*

Description

This example contains p values for a list of SNPs wtih information on chromosome, position and gene symnol.

In the reference below, seven established SNPs are in light blue, 14 new SNPs in dark blue and those failed to replicate in red. The paper size is set to 189 width x 189/2 height (mm) and 1200 dpi resolution. The font is Verdana.

Usage

```
data(hr1420)
```

Format

A data frame

Source

Dr Marcel den Hoed

References

de Hoed M et al. (2013) Heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. Nature Genetics 45(6):621-31, doi: 10.1038/ng.2610.

Examples

```
head(hr1420)
```

---

inf1 *A data containing protein panel*

Description

This data is used to illustrate cis/trans classification, containing the following columns:

Target Target.Short 1 Osteoprotegerin (OPG) OPG 2 C-X-C motif chemokine 11 (CXCL11) CXCL11 3 TNF-related activation cytokine (TRANCE) TRANCE 4 Axin-1 (AXIN1) AXIN1 5 C-C motif chemokine 25 (CCL25) CCL25 6 Tumor necrosis factor (Ligand) superfamily member 12 (TWEAK) TWEAK UniProtID Gene chrom Start End 1 O00300 TNFRSF11B 8 119935796 119964439 2 O14625 CXCL11 4 76954835 76962568 3 O14788 TNFSF11 13 43136872 43182149 4 O15169 AXIN1 16 337440 402673 5 O15444 CCL25 19 8117651 8127534 6 O43508 TNFSF12 17 7452208 7464925

Usage

```
data(inf1)
```

Format

A data frame containing 92 rows and 7 columns

Source

Undisclosed

jma.cojo *A data containing independent GWAS hits as from GCTA*

Description

This data is used to illustrate cis/trans classification, containg the following columns:

```
    prot Chr                SNP       bp refA       freq          b        se
1 4E.BP1  19 chr19:54327313_A_C 54327313    A 0.20550900  0.4510040 0.0243056
2 4E.BP1  19 chr19:54329063_G_T 54329063    T 0.10023500 -0.3233240 0.0333274
3    ADA  19 chr19:54327313_A_C 54327313    A 0.20550900  0.3542660 0.0246266
4    ADA  20 chr20:37456819_C_T 37456819    T 0.00388582 -0.2473080 0.1749800
5    ADA  20 chr20:38196991_G_T 38196991    G 0.00236927 -0.0171435 0.2238980
6    ADA  20 chr20:38603207_A_G 38603207    A 0.17074600 -0.0269075 0.0271976
            p       n  freq_geno         bJ     bJ_se           pJ        LD_r
1 2.48545e-74 6483.69 0.20079500   0.426476 0.0251676  2.07907e-64 -0.13397800
2 4.69307e-22 6480.60 0.08846920  -0.246444 0.0338712  3.44090e-13  0.00000000
3 5.47833e-46 6441.97 0.20079500   0.354266 0.0250171  1.59869e-45  0.00000000
4 1.57618e-01 5553.51 0.00497018  -5.873090 0.2241210 2.32892e-151 -0.00633091
5 9.38970e-01 5556.57 0.00198807 -13.473100 0.3790980 1.18609e-276  0.02467370
6 3.22550e-01 6285.16 0.15009900  -0.299797 0.0278787  5.69806e-27  0.11116200
  UniProtID
1    Q13541
2    Q13541
3    P00813
4    P00813
5    P00813
6    P00813
```

Usage

```
data(jma.cojo)
```

Format

A data frame containing 445 rows and 16 columns

Source

Undisclosed

l51 *An example pedigree data*

Description

The data contains data on 51 individuals in a pedigree. Below it is used for comparing results from various packages.

Usage

```
data(l51)
```

Format

A data frame

Source

Morgan v3.

References

Morgan v3. https://sites.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml

Examples

```
## Not run:
km <- kin.morgan(l51)
k2 <- km$kin.matrix*2

# quantitative trait
library(regress)
r <- regress(qt ~ 1, ~k2, data=l51)
names(r)
r
# qualitative trait
N <- dim(l51)[1]
w <- with(l51,quantile(qt,probs=0.75,na.rm=TRUE))
ped51 <- within(l51, bt <- ifelse(qt<=w,0,1))
d <- regress(bt ~ 1, ~k2, data=ped51)
d
# for other tests not shown here
set.seed(12345)
ped51 <- within(ped51,{r <- rnorm(N); bt[is.na(bt)] <- 0})
library(foreign)
write.dta(ped51,"ped51.dta")

## End(Not run)
```

lukas *An example pedigree*

Description

A multi-generational pedigree containing individual, father, mother IDs and sex.

Usage

```
data(lukas)
```

Format

An example pedigree

Source

Lukas Keller

mao *A study of Parkinson's disease and MAO gene*

Description

The markers are both with actual allele sizes and allele numbers. The dataset is distributed with the GENECOUNTING version 2.0 illustrating gene counting method involving chromosome X. A total of 183 patients and 157 controls (150 males, 190 females) were available, together with five markers in MAOA (monoamine oxidase A) region with alleles 12, 9, 6, 5, 3, and the first three markers were genotyped in all individuals while the fourth and fifth were genotyped for 294 and 304 individuals.

Usage

```
data(mao)
```

Format

A data frame

Source

Dr Helen Latsoudis of Institute of Psychiatry, KCL

References

Zhao JH (2004). 2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis. Bioinformatics 20:1325-1326

meyer *A pedigree data on 282 animals deriving from two generations*

Description

A data frame attributed to Meyer (1989).

"The pedigrees for each of these 282 animals derive from an additional 24 base population (Generation 0) animals that do not have records of their own but, nevertheless, are of interest with respect to the inference on their own additive genetic values. Furthermore, it is presumed that these original 24 base animals are not related to each other. Therefore, the row dimension of u is 306 (282+24)." (Templeman \& Rosa 2004)

Usage

```
data(meyer)
```

Format

A data frame containing 306 records

Source

Meyer K (1989). Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. Genetics, Selection, Evolution 21:317-340.

Tempelman RJ, Rosa GJM. Empirical Bayes Approaches to Mixed Model Inference in Quantitative Genetics. in Saxton AM (Ed). Genetic Analysis of Complex Traits Using SAS, chapter 7. SAS Institute Inc., Cary, NC, USA, 2004

Examples

```
## Not run:
library(gap)
meyer <- within(meyer,{
   g1 <- ifelse(generation==1,1,0)
   g2 <- ifelse(generation==2,1,0)
})
lm(y~-1+g1+g2,data=meyer)
library(MCMCglmm)
m <-MCMCglmm(y~-1+g1+g2,random=animal~1,pedigree=meyer[,1:3],data=meyer,verbose=FALSE)
summary(m)
plot(m)

meyer <- within(meyer,{
   id <- animal
   animal <- ifelse(!is.na(animal),animal,0)
   dam <- ifelse(!is.na(dam),dam,0)
   sire <- ifelse(!is.na(sire),sire,0)
```

```
})
# library(kinship)
# A <- with(meyer,kinship(animal,sire,dam))*2

A <- kin.morgan(meyer)$kin.matrix*2

library(regress)
regress(y~-1+g1+g2,~A,data=meyer)
prior <- list(R=list(V=1, nu=0.002), G=list(G1=list(V=1, nu=0.002)))
m2 <- MCMCgrm(y~-1+g1+g2,prior,meyer,A,singular.ok=TRUE,verbose=FALSE)
summary(m2)
plot(m2)

## End(Not run)
```

`mfblong` *Example data for ACEnucfam*

Description

This is the companion data for ACEnucfam.

Usage

```
data(mfblong)
```

Format

The data is a random subset of the birth weight data from the mental health registry of Norway.

male-a dummy variable for being male; first-a dummy variable for being the first child; midage-a dummy variable for mother aged 20-35 at time of birth; highage-a dummy variable for mother older than 35 at time of birth and birthyr-year of birth minus 1967 (earliest birth year in birth registry).

Source

The data were obtained from the Biometrics website and preprocessed with f.mfb.R.

References

Rabe-Hesketh S, Skrondal A, Gjessing HK. Biometrical modeling of twin and family data using standard mixed model software. Biometrics 2008, 64:280-288

mhtdata *An example data for Manhattan plot with annotation (mhtplot)*

Description

This example contains p values for a list of SNPs whose information regarding chromosome, position and reference seqeuence as with gene annotation is obtained separately.

Usage

```
data(mhtdata)
```

Format

A data frame

Source

Dr Tuomas Kilpelainen at the MRC Epidemiology Unit

References

Kilpelainen1 TO, et al. (2011) Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. Nature Genetics 43(8):753-60, doi: 10.1038/ng.866.

Examples

```
head(mhtdata)
```

nep499 *A study of Alzheimer's disease with eight SNPs and APOE*

Description

This is a study of the neprilysin gene and sporadic Alzheimer's disease in Chinese. There are 257 cases and 242 controls, each with eight SNPs detecting through denaturing high-performance liquid chromatography (DHPLC).

Usage

```
data(nep499)
```

Format

A data frame

Source

Shi J, Zhang S, Tang M, Ma C, Zhao J, Li T, Liu X, Sun Y, Guo Y, Han H, Ma Y, Zhao Z. Mutation Screening and Association Study of the Neprilysin Gene in Sporadic Alzheimer's Disease in Chinese Persons. J Gerontol A: Bio Sci Med Sci 60:301-306, 2005

OPG *An example data for forest plot using METAL output*

Description

This example contains METAL outputs (OPGtbl) as with association statistics from contributing studies (OPGall). It is appropriate to use chr:pos_A1_A2 (A1<=A2) (SNPID) rather than reference id (rsid) due to its variability – therefore a SNPID-rsid mapping file (OPGrsid) is also provided.

Usage

```
data(OPG)
```

Format

Three data frames

Source

SCALLOP consortium

References

pending to give.

Examples

```
data(OPG)
head(OPGtbl)
head(OPGall)
head(OPGrsid)
```

PD *A study of Parkinson's disease and APOE, LRRK2, SNCA makers*

Description

A study of Parkinson's disease and controls with APOE, LRRK2 markers rs10506151, rs10784486, rs1365763, rs1388598, rs1491938, rs1491941 and SNCA markers m770, int4 and SNCA. The column abc indicates if a subject is familial Parkinson's (+), sporadic (-), or controls (Control). Races involved are American Indians (AI), African American (B), and the rest are Caucasians. Diagnosis also included possible (POS), probable (PRO) and definite PDs. AON is the age at onset.

Usage

```
data(PD)
```

Format

A data frame

Source

Prof Abbas Parsian at NIH

References

Parsian et al. ASHG 2005, Toronto

w4 *Results from a GWAS on Chickens*

Description

This example contains p values for a list of SNPs wtih information on chromosome and positions.

Usage

```
data(w4)
```

Format

A data frame

Source

Guo Jun <guojun.yz@gmail.com>

Examples

```
head(w4)
```

Index