Package 'lineup'

January 27, 2015

Version 0.34-1

Date 10/15/2012

Title Lining up two sets of measurements

Author Karl W Broman <kbroman@biostat.wisc.edu>

Maintainer Karl W Broman <kbroman@biostat.wisc.edu>

Depends R (>= 2.10.1), qtl (>= 1.20-15), class, graphics

Description Tools for detecting and correcting sample mix-ups between two sets of measurements, such as between gene expression data on two tissues.

License GPL-3

URL http://www.biostat.wisc.edu/~kbroman/software

Repository CRAN

Date/Publication 2012-10-15 14:26:11

NeedsCompilation yes

R topics documented:

`calc.locallod` *Calculate LOD score at physical position of each gene*

Description

For gene expression data with physical positions of the genes, calculate the LOD score at those positions to assess evidence for local eQTL.

Usage

```
calc.locallod(cross, pheno, pmark, addcovar=NULL, intcovar=NULL,
              verbose=TRUE)
```

Arguments

`cross`	An object of class `"cross"` containing data for a QTL experiment. See the help file for `read.cross` in the R/qtl package (`http://www.rqtl.org`). There must be a phenotype named `"id"` or `"ID"` that contains the individual identifiers.
`pheno`	A data frame of phenotypes (generally gene expression data), stored as individuals x phenotypes. The row names must contain individual identifiers.
`pmark`	Pseudomarkers that are closest to the genes in `pheno`, as output by `find.gene.pseudomarker`.
`addcovar`	Additive covariates passed to `scanone`.
`intcovar`	Interactive covariates passed to `scanone`.
`verbose`	If TRUE, print tracing information.

Details

`cross` and `pheno` must contain exactly the same individuals in the same order. (Use `findCommonID` to line them up.)

We consider the expression phenotypes in batches: those whose closest pseudomarker is the same.

We use Haley-Knott regression to calculate the LOD scores.

Actually, we use a bit of a contortion of the data to force the `scanone` function in R/qtl to calculate the LOD score at a single position.

We omit any transcripts that map to the X chromosome; we can only handle autosomal loci for now.

Value

A vector of LOD scores. The names indicate the gene names (rows in `pheno`).

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

find.gene.pseudomarker, plotEGclass, findCommonID, disteg

Examples

```
## Not run:
##############################
# simulate an eQTL data set
##############################
# genetic map
L <- seq(120, length=8, by=-10)
map <- sim.map(L, n.mar=L/10+1, include.x=FALSE, eq.spacing=TRUE)

# physical map: make all intervals 2x longer
pmap <- rescalemap(map, 2)

# arbitrary locations of 40 local eQTL
thepos <- unlist(map)
theppos <- unlist(pmap)
thechr <- rep(seq(along=map), sapply(map, length))
eqtl.loc <- sort(sample(seq(along=thepos), 40))

x <- sim.cross(map, n.ind=250, type="f2",
               model=cbind(thechr[eqtl.loc], thepos[eqtl.loc], 0, 0))
x$pheno$id <- factor(paste("Mouse", 1:250, sep=""))

# first 20 have eQTL with huge effects
# second 20 have essentially no effect
edata <- cbind((x$qtlgeno[,1:20] - 2)*10+rnorm(prod(dim(x$qtlgeno[,1:20]))),
               (x$qtlgeno[,21:40] - 2)*0.1+rnorm(prod(dim(x$qtlgeno[,21:40]))))
dimnames(edata) <- list(x$pheno$id, paste("e", 1:ncol(edata), sep=""))

# gene locations
theloc <- data.frame(chr=thechr[eqtl.loc], pos=theppos[eqtl.loc])
rownames(theloc) <- colnames(edata)

# mix up 5 individuals in expression data
edata[1:3,] <- edata[c(2,3,1),]
edata[4:5,] <- edata[5:4,]

##############################
# now, the start of the analysis
##############################
x <- calc.genoprob(x, step=1)

# find nearest pseudomarkers
pmark <- find.gene.pseudomarker(x, pmap, theloc, "prob")

# calculate LOD score for local eQTL
locallod <- calc.locallod(x, edata, pmark)

# take those with LOD > 100 [which will be the first 20]
```

```
edatasub <- edata[,locallod>100,drop=FALSE]

# calculate distance between individuals
#     (prop'n mismatches between obs and inferred eQTL geno)
d <- disteg(x, edatasub, pmark)

# plot distances
plot(d)

# summary of apparent mix-ups
summary(d)

# plot of classifier for first eQTL
plotEGclass(d)

## End(Not run)
```

`combinedist` *Combine distance matrices into a single such*

Description

Combine multiple distance matrices into a single distance matrix providing an overall summary

Usage

```
combinedist(..., method=c("median", "mean"))
```

Arguments

`...`	Set of distance matrices, as calculated by `distee` or `disteg`.
`method`	Indicates whether to summarize using the median or the mean.

Details

The row and column names of the input distance matrices define the individual IDs.

If the input distance matrices all have an attribute `"denom"` (for denominator) and `method="mean"`, we use a weighted mean, weighted by the denominators. This could be used to calculate an overall proportion.

Value

A distance matrix, with class `"lineupdist"`. The individual IDs are in the row and column names.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

distee, disteg, summary.lineupdist

Examples

```
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create three data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
w <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows of x
x[51:53,] <- x[c(52,53,51),]

# add column and row names
dimnames(x) <- dimnames(y) <- dimnames(w) <-
   list(paste("ind", 1:100, sep=""), paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and of the other two matrices
corxy <- corbetw2mat(x, y)
corxw <- corbetw2mat(x, w)

# using columns with corr > 0.75,
# calculate distance (using "correlation" as a measure...really similarity)
dxy <- distee(x[,corxy>0.75], y[,corxy>0.75], d.method="cor", labels=c("x", "y"))
dxw <- distee(x[,corxw>0.75], w[,corxw>0.75], d.method="cor", labels=c("x", "w"))

d <- combinedist(dxy, dxw)

summary(d)
```

---

corbetw2mat *Calculate correlations between columns of two matrices*

Description

For matrices x and y, calculate the correlation between columns of x and columns of y.

Usage

```
corbetw2mat(x, y, what=c("paired", "bestright", "bestpairs", "all"),
            corthresh=0.9)
```

Arguments

x	A numeric matrix.
y	A numeric matrix with the same number of rows as x.
what	Indicates which correlations to calculate and return. See value, below.
corthresh	Threshold on correlations if what="bestpairs".

Details

Missing values (NA) are ignored, and we calculate the correlation using all complete pairs, as in cor with use="pairwise.complete.obs".

Value

If what="paired", the return value is a vector of correlations, between columns of x and the corresponding column of y. x and y must have the same number of columns.

If what="bestright", we return a data frame of size ncol(x) by 3, with the ith row being the maximum correlation between column i of x and a column of y, and then the y-column index and y-column name with that correlation. (In case of ties, we give the first one.)

If what="bestpairs", we return a data frame with five columns, containing all pairs of columns (with one in x and one in y) with correlation $\geq$ corthresh. Each row corresponds to a column pair, and contains the correlation and then the x- and y-column indices followed by the x- and y-column names.

If what="all", the output is a matrix of size ncol(x) by ncol(y), with all correlations between columns of x and columns of y.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

distee, findCommonID

Examples

```
# a variance matrix
V <- diag(rep(0.5, 5)) + 0.5
D <- chol(V)

# simulate two correlated matrices
x <- matrix(rnorm(100), ncol=5)
y <- matrix(rnorm(100), ncol=5)

# create shuffled version of the second matrix
u <- sample(1:ncol(y))
z <- y[,u]

# correlations with paired columns
```

```
corbetw2mat(x, y)

# the same with y columns shuffled
corbetw2mat(x, z)

# for each column x, find column of y with max correlation
corbetw2mat(x, y, what="bestright")

# the same with y columns shuffled
corbetw2mat(x, z, what="bestright")


# all pairs of columns with correlation >= 0.6
corbetw2mat(x, y, what="bestpairs", corthresh=0.6)

# the same with y columns shuffled
corbetw2mat(x, z, what="bestpairs", corthresh=0.6)

# all correlations
corbetw2mat(x, y, what="all")
```

`distee` *Calculate distance between two gene expression data sets*

Description

Calculate a distance between all pairs of individuals for two gene expression data sets

Usage

```
distee(e1, e2, d.method=c("rmsd", "cor"), labels=c("e1","e2"),
       verbose=TRUE)
```

Arguments

`e1`	Numeric matrix of gene expression data, as individuals x genes. The row and column names must contain individual and gene identifiers.
`e2`	(Optional) Like `e1`. An appreciable number of individuals and genes must be in common.
`d.method`	Calculate inter-individual distance as RMS difference or as correlation.
`labels`	Two character strings, to use as labels for the two data matrices in subsequent output.
`verbose`	if TRUE, give verbose output.

Details

We calculate the pairwise distance between all individuals (rows) in `e1` and all individuals in `e2`. This distance is either the RMS difference (`d.method="rmsd"`) or the correlation (`d.method="cor"`).

Value

A matrix with `nrow(e1)` rows and `nrow(e2)` columns, containing the distances. The individual IDs are in the row and column names. The matrix is assigned class `"lineupdist"`.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

pulldiag, omitdiag, summary.lineupdist, plot2dist, disteg, corbetw2mat

Examples

```
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
summary(d1)
summary(d2)

# order to put matches together
```

```
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)

# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)
```

---

disteg *Calculate distance between two gene expression data sets*

Description

Calculate a distance between all pairs of individuals for two gene expression data sets

Usage

```
disteg(cross, pheno, pmark, min.genoprob=0.99,
       k=20, min.classprob=0.8, classprob2drop=1, repeatKNN=TRUE,
       max.selfd=0.3, phenolabel="phenotype",
       weightByLinkage=FALSE,
       map.function=c("haldane", "kosambi", "c-f", "morgan"),
       verbose=TRUE)
```

Arguments

cross	An object of class "cross" containing data for a QTL experiment. See the help file for read.cross in the R/qtl package (http://www.rqtl.org). There must be a phenotype named "id" or "ID" that contains the individual identifiers.
pheno	A data frame of phenotypes (generally gene expression data), stored as individuals x phenotypes. The row names must contain individual identifiers.
pmark	Pseudomarkers that are closest to the genes in pheno, as output by find.gene.pseudomarker.
min.genoprob	Threshold on genotype probabilities; if maximum probability is less than this, observed genotype taken as NA.
k	Number of nearest neighbors to consider in forming a k-nearest neighbor classifier.
min.classprob	Minimum proportion of neighbors with a common class to make a class prediction.
classprob2drop	If an individual is inferred to have a genotype mismatch with classprob > this value, treat as an outlier and drop from the analysis and then repeat the KNN construction without it.
repeatKNN	If TRUE, repeat k-nearest neighbor a second time, after omitting individuals who seem to not be self-self matches

max.selfd	Min distance from self (as proportion of mismatches between observed and predicted eQTL genotypes) to be excluded from the second round of k-nearest neighbor.
phenolabel	Label for expression phenotypes to place in the output distance matrix.
weightByLinkage	If TRUE, weight the eQTL to account for their relative positions (for example, two tightly linked eQTL would each count about 1/2 of an isolated eQTL)
map.function	Used if weightByLinkage is TRUE
verbose	if TRUE, give verbose output.

Details

We consider the expression phenotypes in batches, by which pseudomarker they are closest to. For each batch, we pull the genotype probabilities at the corresponding pseudomarker and use the individuals that are in common between cross and pheno and whose maximum genotype probability is above min.genoprob, to form a classifier of eQTL genotype from expression values, using k-nearest neighbor (the function knn). The classifier is applied to all individuals with expression data, to give a predicted eQTL genotype. (If the proportion of the k nearest neighbors with a common class is less than min.classprob, the predicted eQTL genotype is left as NA.)

If repeatKNN is TRUE, we repeat the construction of the k-nearest neighbor classifier after first omitting individuals whose proportion of mismatches between observed and inferred eQTL genotypes is greater than max.selfd.

Finally, we calculate the distance between the observed eQTL genotypes for each individual in cross and the inferred eQTL genotypes for each individual in pheno, as the proportion of mismatches between the observed and inferred eQTL genotypes.

If weightByLinkage is TRUE, we use weights on the mismatch proportions for the various eQTL, taking into account their linkage. Two tightly linked eQTL will each be given half the weight of a single isolated eQTL.

Value

A matrix with nind(cross) rows and nrow(pheno) columns, containing the distances. The individual IDs are in the row and column names. The matrix is assigned class "lineupdist".

The names of the genes that were used to construct the classifier are saved in an attribute "retained".

The observed and inferred eQTL genotypes are saved as attributes "obsg" and "infg".

The denominators of the proportions that form the inter-individual distances are in the attribute "denom".

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

distee, summary.lineupdist, pulldiag, omitdiag, findCommonID, find.gene.pseudomarker, calc.locallod, plot.lineupdist, knn, plotEGclass

Examples

```
##############################
# simulate an eQTL data set
##############################
# genetic map
L <- seq(120, length=8, by=-10)
map <- sim.map(L, n.mar=L/10+1, include.x=FALSE, eq.spacing=TRUE)

# physical map: make all intervals 2x longer
pmap <- rescalemap(map, 2)

# arbitrary locations of 40 local eQTL
thepos <- unlist(map)
theppos <- unlist(pmap)
thechr <- rep(seq(along=map), sapply(map, length))
eqtl.loc <- sort(sample(seq(along=thepos), 40))

x <- sim.cross(map, n.ind=250, type="f2",
               model=cbind(thechr[eqtl.loc], thepos[eqtl.loc], 0, 0))
x$pheno$id <- factor(paste("Mouse", 1:250, sep=""))

# first 20 have eQTL with huge effects
# second 20 have essentially no effect
edata <- cbind((x$qtlgeno[,1:20] - 2)*10+rnorm(prod(dim(x$qtlgeno[,1:20]))),
               (x$qtlgeno[,21:40] - 2)*0.1+rnorm(prod(dim(x$qtlgeno[,21:40]))))
dimnames(edata) <- list(x$pheno$id, paste("e", 1:ncol(edata), sep=""))

# gene locations
theloc <- data.frame(chr=thechr[eqtl.loc], pos=theppos[eqtl.loc])
rownames(theloc) <- colnames(edata)

# mix up 5 individuals in expression data
edata[1:3,] <- edata[c(2,3,1),]
edata[4:5,] <- edata[5:4,]

##############################
# now, the start of the analysis
##############################
x <- calc.genoprob(x, step=1)

# find nearest pseudomarkers
pmark <- find.gene.pseudomarker(x, pmap, theloc, "prob")

# calculate LOD score for local eQTL
locallod <- calc.locallod(x, edata, pmark)

# take those with LOD > 100 [which will be the first 20]
edatasub <- edata[,locallod>100,drop=FALSE]

# calculate distance between individuals
#     (prop'n mismatches between obs and inferred eQTL geno)
d <- disteg(x, edatasub, pmark)
```

```
# plot distances
plot(d)

# summary of apparent mix-ups
summary(d)

# plot of classifier for first eQTL
plotEGclass(d)
```

`find.gene.pseudomarker` *Find nearest peudomarker to each gene*

Description

Pull out the pseudomarker that is closest to the position of each of a series of genes.

Usage

```
find.gene.pseudomarker(cross, pmap, geneloc, where=c("prob", "draws"))
```

Arguments

`cross`	An object of class `"cross"` containing data for a QTL experiment. See the help file for `read.cross` in the R/qtl package (`http://www.rqtl.org`).
`pmap`	A physical map of the markers in `cross`, with locations in Mbp. This is a list whose components are the marker locations on each chromosome.
`geneloc`	A data frame specifying the physical locations of the genes. There should be two columns, `chr` for chromosome and `pos` for position in Mbp. The rownames should indicate the gene names.
`where`	Indicates whether to pull pseudomarkers from the genotype probabilities (produced by `calc.genoprob`) or from the imputed genotypes (produced by `sim.geno`).

Details

We first convert positions (by interpolation) from those contained within `cross` to physical coordinates contained in `pmap`. We then use `find.pseudomarker` to identify the closest pseudomarker to each gene location.

We also include the positions of the pseudomarkers, and we print a warning message if pseudomarkers are > 2 Mbp from the respective gene.

Value

A data frame with columns `chr` (the chromosome) and `pmark` (the name of the pseudomarker). The third column `pos` contains the Mbp position of the pseudomarker. The final column is the signed distance between the gene and the pseudomarker. The rownames indicate the gene names.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

find.pseudomarker, find.pseudomarkerpos, plotEGclass, disteg, calc.locallod

Examples

```
## Not run:
##############################
# simulate an eQTL data set
##############################
# genetic map
L <- seq(120, length=8, by=-10)
map <- sim.map(L, n.mar=L/10+1, include.x=FALSE, eq.spacing=TRUE)

# physical map: make all intervals 2x longer
pmap <- rescalemap(map, 2)

# arbitrary locations of 40 local eQTL
thepos <- unlist(map)
theppos <- unlist(pmap)
thechr <- rep(seq(along=map), sapply(map, length))
eqtl.loc <- sort(sample(seq(along=thepos), 40))

x <- sim.cross(map, n.ind=250, type="f2",
               model=cbind(thechr[eqtl.loc], thepos[eqtl.loc], 0, 0))
x$pheno$id <- factor(paste("Mouse", 1:250, sep=""))

# first 20 have eQTL with huge effects
# second 20 have essentially no effect
edata <- cbind((x$qtlgeno[,1:20] - 2)*10+rnorm(prod(dim(x$qtlgeno[,1:20]))),
               (x$qtlgeno[,21:40] - 2)*0.1+rnorm(prod(dim(x$qtlgeno[,21:40]))))
dimnames(edata) <- list(x$pheno$id, paste("e", 1:ncol(edata), sep=""))

# gene locations
theloc <- data.frame(chr=thechr[eqtl.loc], pos=theppos[eqtl.loc])
rownames(theloc) <- colnames(edata)

# mix up 5 individuals in expression data
edata[1:3,] <- edata[c(2,3,1),]
edata[4:5,] <- edata[5:4,]

##############################
# now, the start of the analysis
##############################
x <- calc.genoprob(x, step=1)

# find nearest pseudomarkers
pmark <- find.gene.pseudomarker(x, pmap, theloc, "prob")
```

```
# calculate LOD score for local eQTL
locallod <- calc.locallod(x, edata, pmark)

# take those with LOD > 100 [which will be the first 20]
edatasub <- edata[,locallod>100,drop=FALSE]

# calculate distance between individuals
#     (prop'n mismatches between obs and inferred eQTL geno)
d <- disteg(x, edatasub, pmark)

# plot distances
plot(d)

# summary of apparent mix-ups
summary(d)

# plot of classifier for first eQTL
plotEGclass(d)

## End(Not run)
```

`findCommonID` *Find individuals in common between a cross and a phenotype matrix*

Description

Identify which individuals are in common between a QTL mapping data set and a matrix of phenotypes, series of genes.

Usage

```
findCommonID(id1, id2)
```

Arguments

`id1`	A character vector of individual IDs. This can also be a QTL cross object (see `read.cross`), in which case `getid` is used to grab individual IDs, or a matrix or data frame, in which case the rownames are taken to be IDs.
`id2`	Like `id1`, can be a character vector, a cross or a matrix/data frame.

Value

A list with three components:

First, a data frame with rows corresponding to all individuals (across the two sets of individual IDs) and three columns: `indexInFirst` and `indexInSecond` contain numeric indices to the locations of the individuals within `cross` and `pheno`, and `inBoth` is a logical vector to indicate which individuals appear in both crosses. The row names are the individual identifiers.

The second and third components are vectors of indices in `id1` and `id2`, respectively, indicating the paired locations of the individuals that are in common.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

calc.locallod, corbetw2mat

Examples

```
id1 <- sample(LETTERS[1:5])
id2 <- LETTERS[3:8]
findCommonID(id1, id2)

x <- matrix(0, nrow=length(id2), ncol=3)
rownames(x) <- id2
findCommonID(id1, x)
```

---

fscale *Standardize the columns of a matrix*

Description

Standardize each column in a matrix, so that the columns have mean 0 and SD 1.

Usage

```
fscale(x)
```

Arguments

x	A numeric matrix.

Details

Missing values (NA) are ignored and left as is.

If there is just 1 non-missing value in a column, it is left as is.

This function uses a one-pass algorithm to calculate the mean and SD, which is fast but can show a bit of round-off error.

Value

A matrix of the same form as the input, but with columns transformed to have mean 0 and SD 1.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

scale

Examples

```
x <- matrix(1:10, ncol=2)
y <- fscale(x)
```

---

lineupversion *Installed version of R/lineup*

Description

Print the version number of the currently installed version of R/lineup.

Usage

```
lineupversion()
```

Value

A character string with the version number of the currently installed version of R/lineup.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

Examples

```
lineupversion()
```

---

omitdiag *Replace the diagonal in a distance matrix with missing values*

Description

Replace the diagonal (that is, self-self distances) from a distance matrix calculated by distee or disteg with missing values (so that only self-nonself distances are left).

Usage

```
omitdiag(d)
```

Arguments

d	A distance matrix calculated by distee or disteg.

Details

We use the row and column names to identify which entries are self-self.

Value

A matrix of the same form as the input, but with self-self distances replaced with NA.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

pulldiag, distee, disteg, summary.lineupdist, plot2dist, plot.lineupdist

Examples

```
## Not run:
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
```

```
summary(d1)
summary(d2)

# order to put matches together
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)

# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)

## End(Not run)
```

`plot.lineupdist` *Plot summary of inter-individual distances*

Description

Plot histograms of self-self and self-nonself distances from a distance matrix calculated by `distee` or `disteg`.

Usage

```
## S3 method for class 'lineupdist'
plot(x, breaks, add.rug=TRUE, what=c("both", "ss", "sn"), ...)
```

Arguments

`x`	Output of `distee` or `disteg`.
`breaks`	Optional vector of breaks, passed to `hist`, though if it is length 1, we interpret it as the number of breaks and ensure that both histograms use the same set of breaks.
`add.rug`	If true, also include `rug` below histograms.
`what`	Indicates whether to plot both self-self and self-nonself distances (or correlations) or just one or the other. (`"ss"` indicates self-self and `"sn"` indicates self-nonself.)
`...`	Ignored at this point.

Details

We call `pulldiag` and `omitdiag` to get the self-self and self-nonself distances.

If all of the self-self distances are missing, we plot just the self-nonself distances.

Value

None.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

pulldiag, distee, plot2dist

Examples

```
## Not run:
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
summary(d1)
summary(d2)

# order to put matches together
```

```
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)

# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)

## End(Not run)
```

`plot2dist` *Plot two sets of inter-individual distances against one another*

Description

Plot two sets of inter-individual distances against one another, colored by self and non-self distances.

Usage

```
plot2dist(d1, d2, hirow, hicol, xlab, ylab, smoothScatter=FALSE,
       colself="black", colnonself="gray", colhirow="green", colhicol="orange", ...)
```

Arguments

`d1`	Output of `distee`.
`d2`	Output of `distee`.
`hirow`	Names of rows to highlight in green.
`hicol`	Names of columns to highlight in orange.
`xlab`	X-axis label (optional)
`ylab`	Y-axis label (optional)
`smoothScatter`	If TRUE, plot non-self distances with `smoothScatter`; if FALSE, use `plot`.
`colself`	Color to use for the self-self points. If NULL, these aren't plotted.
`colnonself`	Color to use for the non-self points. If NULL, these aren't plotted.
`colhirow`	Color to use for the `hirow` points. If NULL, these aren't plotted.
`colhicol`	Color to use for the `hicol` points. If NULL, these aren't plotted.
`...`	Passed to `plot` and `points`.

Value

None.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

pulldiag, distee, summary.lineupdist

Examples

```
## Not run:
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
summary(d1)
summary(d2)

# order to put matches together
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)
```

```
# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)

## End(Not run)
```

`plotEGclass` *Plot classifier of eQTL genotype from expression data*

Description

Diagnostic plot of one of the eQTL classifiers from the results of `disteg`: generally expression phenotype against observed eQTL genotype, colored by inferred eQTL genotype.

Usage

```
plotEGclass(d, eqtl=1, outercol="inferred", innercol="observed",
            thecolors=c("blue","green","red","orange"), ...)
```

Arguments

`d`	Output of `disteg`.
`eqtl`	Numeric index or a character vector (of the form "1@102.35") indicating the eQTL to consider.
`outercol`	Indicates how to color the outer edge of the points: `"observed"` indicates to color based on observed genotypes; `"inferred"` indicates to color based on inferred genotypes; otherwise, give a color.
`innercol`	Like `outercol`, but indicating the interior of the points.
`thecolors`	The colors to use in the plot. The last element (after the number of genotypes) indicates the color to use for missing values.
`...`	Passed to `plot` and `points`.

Details

The function produces a diagnostic plot for studying one of the k-nearest neighbor classifiers underlying the output from `disteg`.

In the case of one expression phenotype attached to the selected eQTL, the plot is a dot plot of gene expression against observed eQTL genotype.

In the case of two expression phenotypes, the plot is a scatterplot of the two expression phenotypes against each other.

In the case of more than two expression phenotypes, we use `pairs` to produce a matrix of scatterplots.

Value

None.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

disteg, plot.lineupdist, plot2dist, knn

Examples

```
## Not run:
##############################
# simulate an eQTL data set
##############################
# genetic map
L <- seq(120, length=8, by=-10)
map <- sim.map(L, n.mar=L/10+1, include.x=FALSE, eq.spacing=TRUE)

# physical map: make all intervals 2x longer
pmap <- rescalemap(map, 2)

# arbitrary locations of 40 local eQTL
thepos <- unlist(map)
theppos <- unlist(pmap)
thechr <- rep(seq(along=map), sapply(map, length))
eqtl.loc <- sort(sample(seq(along=thepos), 40))

x <- sim.cross(map, n.ind=250, type="f2",
               model=cbind(thechr[eqtl.loc], thepos[eqtl.loc], 0, 0))
x$pheno$id <- factor(paste("Mouse", 1:250, sep=""))

# first 20 have eQTL with huge effects
# second 20 have essentially no effect
edata <- cbind((x$qtlgeno[,1:20] - 2)*10+rnorm(prod(dim(x$qtlgeno[,1:20]))),
               (x$qtlgeno[,21:40] - 2)*0.1+rnorm(prod(dim(x$qtlgeno[,21:40]))))
dimnames(edata) <- list(x$pheno$id, paste("e", 1:ncol(edata), sep=""))

# gene locations
theloc <- data.frame(chr=thechr[eqtl.loc], pos=theppos[eqtl.loc])
rownames(theloc) <- colnames(edata)

# mix up 5 individuals in expression data
edata[1:3,] <- edata[c(2,3,1),]
edata[4:5,] <- edata[5:4,]

##############################
# now, the start of the analysis
##############################
x <- calc.genoprob(x, step=1)
```

```
# find nearest pseudomarkers
pmark <- find.gene.pseudomarker(x, pmap, theloc, "prob")

# calculate LOD score for local eQTL
locallod <- calc.locallod(x, edata, pmark)

# take those with LOD > 100 [which will be the first 20]
edatasub <- edata[,locallod>100,drop=FALSE]

# calculate distance between individuals
#     (prop'n mismatches between obs and inferred eQTL geno)
d <- disteg(x, edatasub, pmark)

# plot distances
plot(d)

# summary of apparent mix-ups
summary(d)

# plot of classifier for first eQTL
plotEGclass(d)

## End(Not run)
```

`pulldiag` *Pull out the diagonal from a distance matrix*

Description

Pull out the diagonal from a distance matrix calculated by `distee` (that is, self-self distances).

Usage

```
pulldiag(d)
```

Arguments

d	A distance matrix calculated by `distee`.

Details

We use the row and column names to identify which entries are self-self.

Value

A vector with the self-self distances.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

omitdiag, distee, disteg, summary.lineupdist, plot2dist, plot.lineupdist

Examples

```
## Not run:
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
summary(d1)
summary(d2)

# order to put matches together
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)
```

```
# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)

## End(Not run)
```

`subset.lineupdist` *Subsetting distance matrix*

Description

Pull out a specified set of rows and columns from a distance matrix calculated by `distee` or `disteg`.

Usage

```
## S3 method for class 'lineupdist'
subset(x, rows, cols, ...)
## S3 method for class 'lineupdist'
x[rows, cols]
```

Arguments

`x`	A distance matrix object as obtained from `distee` or `disteg`.
`rows`	Optional vector of selected rows.
`cols`	Optional vector of selected columns.
`...`	Ignored at this point.

Value

The input distance matrix object, but with only the specified subset of the data.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

`disteg`, `distee`, `pulldiag`

summary.lineupdist *Summarize inter-individual distances*

Description

Summarize the results of distee or disteg, with inter-individual distances between two sets of gene expression data.

Usage

```
## S3 method for class 'lineupdist'
summary(object, cutoff, dropmatches=TRUE,
                  reorder=c("bydistance", "alignmatches", "no"), ...)
```

Arguments

object	Output of distee or disteg.
cutoff	(Optional) Cutoff on correlation/distance, with rows in the results only being kept if the best distance/correlation is above this cutoff or the self-self result is not missing and is above this cutoff.
dropmatches	If TRUE, omit rows for which an individual's best match is itself.
reorder	If "bydistance", reorder rows by increasing distance (or decreasing correlation) to the best match and then by decreasing distance (or decreasing correlation) to self; if "alignmatches", group related errors together; if "no", leave as is.
...	Passed to print.data.frame.

Value

A list with two components: the distances summarized by row and the distances summarized by column.

For each individual, we calculate the minimum distance to others, next-smallest distance, the self-self distance, the mean and SD of the distances to others, and finally indicate the individual (or individuals) that is closest.

Author(s)

Karl W Broman, <kbroman@biostat.wisc.edu>

See Also

pulldiag, omitdiag, distee, disteg, plot2dist, plot.lineupdist

Examples

```
## Not run:
# simulate MVN, 100 individuals, 40 measurements (of which 20 are just noise)
V <- matrix(0.3, ncol=20, nrow=20) + diag(rep(0.5, 20))
D <- chol(V)
z <- matrix(rnorm(20*100), ncol=20)

# create two data matrices as z + noise
x <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))
y <- cbind(z + rnorm(20*100, 0, 0.2), matrix(rnorm(20*100), ncol=20))

# permute some rows
x[51:53,] <- x[c(52,53,51),]
y[41:42,] <- y[42:41,]

# add column and row names
dimnames(x) <- dimnames(y) <- list(paste("ind", 1:100, sep=""),
                                   paste("gene", 1:40, sep=""))

# calculate correlations between cols of x and cols of y
thecor <- corbetw2mat(x, y)

# subset x and y, taking only columns with corr > 0.75
xs <- x[,thecor > 0.8]
ys <- y[,thecor > 0.8]

# calculate distance (using "RMS difference" as a measure)
d1 <- distee(xs, ys, d.method="rmsd", labels=c("x","y"))

# calculate distance (using "correlation" as a measure...really similarity)
d2 <- distee(xs, ys, d.method="cor", labels=c("x", "y"))

# pull out the smallest 8 self-self correlations
sort(pulldiag(d2))[1:8]

# summary of results
summary(d1)
summary(d2)

# order to put matches together
summary(d2, reorder="alignmatches")

# plot histograms of RMS distances
plot(d1)

# plot histograms of correlations
plot(d2)

# plot distances against one another
plot2dist(d1, d2)

## End(Not run)
```

Index