# Package 'hegp'

June 2, 2020

**Type** Package

**Title** Homomorphic Encryption of Genotypes and Phenotypes

**Version** 1.0.0

**Imports** parallel, mixed.model.gwas

**Author** Richard Mott

**Maintainer** The package maintainer <r.mott@ucl.ac.uk>

**Description** Uses random orthogonal matrices to homomorphically encrypt phenotypes and genotypes for quantitative genetic analysis.

**License** GPL3.0

**Encoding** UTF-8

**LazyData** true

## R topics documented:



---

.HEGP *Homomorphic Encryption of Genotypes and Phenotypes*

---

### Description

Overview of the use of orthogonal matrix keeys for homomorphic encryption of quantitative genetic data

## Details

This package implements a method of homomorphic encryption of genotypes and phenotypes suitable for quantitative genetic analysis. Full details are in the paper (Mott et al Genetics, 2020). Briefly, the encryption key is a randomly generated orthogonal matrix that is multiplied into the plaintext phenotype, genotypes and optionally any covariates to produce a ciphertext that closely resembles samples from a Normal distribution. The othogonal transformation leaves the log-likelihood of the data unchanged and hence all inferences about parameter estimates, heritability and p-values of association are unchanged by the transformation.

Specifically, suppose we have a standardised vector $y$ of $n$ phenotypes and an $n \times p$ matrix $G$ of standardised SNP dosages. Suppose further the phenotypic variance-covariance matrix $V$ is modelled as

$$V = K\sigma_g^2 + I\sigma_e^2$$

where $K = GG^T/p$ is the additive genetic relationship matrix (sometimes called a kinship matrix). Then the mixed model

$$y = X\beta + e$$

describes the relationship betwen the phenotype and a set of fixed effects, represented by the $n \times k$ design matrix $X$, one of which will be a SNP (a column of $G$), others may be covariates such as sex. The residual vector $e$ has the same variance matrix $V$ as $y$. The variance matrix $V$ can be decomposed into its matrix square root via its eigen decomposition:

$$V = E^T \Lambda E = (E^T \Lambda^{0.5} E)(E^T \Lambda^{0.5} E) = A^2$$

where $E$ is the orthogonal matrix of eigenvectors and $\Lambda$ a diagonal matrix of positive eigenvalues. Then the mixed model transformation

$$A^{-1}y = A^{-1}X\beta + A^{-1}e$$

converts the mixed model to ordinary least squares in which the variance matrix of the error is the identity matrix.

Now suppose $P$ is an $n \times n$ orthogonal matrix (ie so $PP^T = I$. Then if we replace the plaintext

$$D = [y, G, X, V]$$

by the ciphertext

$$D(P) = [Py, PG, PX, PVP^T]$$

the resulting ciphertext mixed model has the same log-likelihood, and the maximum likelihood estimators of the fixed effects $\beta$ are unchanged, as are the p-values of genetic association and the estimates of the variance components $\sigma_g^2, \sigma_e^2$, the heritability $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$ and the linkage disequilibrium between SNPs. Thus the linear mixed model is invariant under orthogonal transformation.

Furthermore, if we sample the orthogonal transformation at random from the Stiefel Manifold then the transformed ciphertext $D(P)$ closely resemble a sample of Gaussian deviates, and can be thought of as a homomorphic encryption of the plaintext $D$, that preserves all the essential characteristics of the data while obscuring the genotypes of the individuals. The Stiefel Manifold samples orthogonal matrices as follows:

(i) Simulate an $n \times n$ matrix $M$ whose entries are all iid N(0,1).

(ii) Compute the eigen-decomposition of the symmetric matrix $M^T M = Q^T SQ$ where $Q$ is $n \times n$ orthogonal and $S$ is diagonal with positive entries.

(iii) Return the orthogonal matrix $P = MQ^T S^{-0.5} Q$ where $S^{-0.5}$ is the diagonal matrix whose elements are the reciprocals of the square roots of the eigenvalues.

The HEGP package implements this scheme. It encapsulates the phenotype, genotypes, covariates and any ancilliary files such as the genetic relationship matrix and the map of SNP coordinates as a special dataset $D$, implemented as a list of conformant R objects using the function `build.D()`. These Datasets can then be manipulated by the other functions in the package. For instance orthogonal encryption is acocmplished by `encrypt.D`. Functions to sample random orthogonal matrices are provided in `make.encrypter`, `rsteifel` (based on the Rsteifel package), and functions to perform mixed model GWAS `basic.mm.gwas` and simple GWAS `basic.gwas` are also provided. There is also a function `qnorm.D` to add an addition level of security by replacing each column of encrypted data by quantiles from a Gaussian distribution.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

---

`basic.gwas` *basic.gwas*

---

## Description

Perform a standard genome wide association analysis. Used to check that plaintext and ciphertext data produce the same gwas results.

## Usage

```
basic.gwas( D, mc.cores=10 )
```

## Arguments

| | |
|---|---|
| `D` | A Dataset, as generated by `build.D` or `encrypt.D` |
| `mc.cores` | Number of cores for parellelisation |

## Details

Each vector of SNP dosages is tested for assocation with the phenotype by simple linear regression (i.e. no mixed model).

## Value

A dataframe containing the logP of each tested SNP joined to the columns of D$map

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

## See Also

build.D encrypt.D basic.mm.gwas

---

basic.mm.gwas *Mixed Model GWAS*

---

## Description

Perform a mixed-model GWAS to check that plaintext and ciphertext produce the same results.

## Usage

```
basic.mm.gwas(D, mc.cores=10)
```

## Arguments

| | |
|---|---|
| D | A Dataset as generated by build.D or encrypt.D |
| mc.cores | Number of cores over which to parallelize computation |

## Details

A standard mixed model is fitted to the data, using a SNP-based genetic reelationship matrix. The phenotype and genotype are then transformed and each transformed SNP is tested for association with the transformed phenotype. Uses the function mixed.model.gwas from the package mixed.model.gwas

## Value

A dataframe containing the logP of each tested SNP joined to the columns of D$map

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

## See Also

build.D encrypt.D basic.gwas

build.D *build.D*

## Description

Create a Dataset object from its constituent components, namely the phenotype, genotype dosages, covariates, and optionally the physical map and kinship matrix. Once created, such a dataset can be manipulated by other functions such as encrypt.D, link{basic.gwas}, basic.mm.gwas

## Usage

```
build.D( y, dosages, cov=NULL, map=NULL, kinship=FALSE )
```

## Arguments

| | |
|---|---|
| y | Numeric phenotype vector |
| dosages | Matrix of genotype dosages |
| cov | Optional matrix of covariates |
| map | Optional data frame of information about genotypes. If supplied, the i'th row of map refers to the i'th column of the genotype dosages. |
| kinship | Optional switch to generate a genetic relationship matrix from the genotype dosages |

## Value

A list with the components y=y.s, geno=geno, cov=cov, map=map, maf=af

| | |
|---|---|
| y | vector of phenotypes, scaled to have zero mean and variance equal to one |
| geno | matrix of genotype dosages, each column (SNP) scaled to have zero mean and variance equal to one |
| cov | matrix of covariates. If the input covariate matrix is NULL this is a vector of ones |
| map | optional dataframe of information about SNPs, e.g. chromosome and base-pair coordinate |
| af | allele frequencies of the SNPs, computed from the genotype dosages |
| kinship | optional genetic relationship matrix |

## Note

No missing values are allowed. The dimensions of the phenotypes and genotypes are made compatible by matching the rownames of the genotypes with the names of the phenotypes. If a genetic relationship matrix is calculated it uses the function make.kinship from the library mixed.model.gwas.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

## See Also

encrypt.D basic.gwas basic.mm.gwas

## Examples

```
#
```

---

encrypt.D *encrypt.D*

---

## Description

Encrypt or decrypt a dataset.

## Usage

```
encrypt.D( D, encrypter, invert=FALSE, kinship=FALSE )
```

## Arguments

| | |
|---|---|
| D | A Dataset to be encrypted, generateed by build.D |
| encrypter | An encrypter as generated by make.encrypter |
| invert | Decrypt the data by using the inverse encrypter (matrix transpose) |
| kinship | An optional Boolean switch, which if TRUE will create an encrypted kinship matrix of dimension $N \times N$, where $N$ is the number of individuals. |

## Value

An encrypted (ciphertext) or decrypted (plaintext) Dataset derived from the input data by applying thee encryptor to it

## Note

The kinship (genetic relationship) matrix is defined as $K = GG^T/p$ where $G$ is the standardised matrix of SNP genotype dosages and $p$ the number of SNPs. The function make.kinship in the package mixed.model.gwas is used for this purpose.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

## See Also

build.D make.encrypter

`make.encrypter` *Create encryption keys for a Dataset*

## Description

Sample a series of orthogonal matrices suitable for encrypting a given Dataset object.

## Usage

```
make.encrypter( D, blocksize=0 )
```

## Arguments

| | |
|---|---|
| `D` | A Dataset object, created by `build.D` |
| `blocksize` | Optional size of encryption blocks. Each block of individuals is encrypted separately. If blocksize is zero then a single encrypter is generated. |

## Details

Create random orthogonal encryption keys for the dataset D created by a call to `build.D`. Each encryption key is a random orthogonal matrix generated from the Steifel manifold. If the dataset contains $N$ individuals then if $blocksize > 0$, $N/blocksize + 1$ keys are generated. Most keys are of dimension $blocksize * blocksize$ with the final key with smaller dimension to make the sum of the dimensions of the keys equal to $N$. The Dataset can then be encrypted using the orthogonal keys by `encrypt.D`.

## Value

A list with elements

| | |
|---|---|
| `blocks` | The number of blocks |
| `block` | a list of encryption keys, each a matrix |

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

## See Also

`build.D` `encrypt.D`

qnorm.D *qnorm.D*

## Description

Replace a phenotype and each vector of genotype dosages by their Normal quantiles

## Usage

```
qnorm.D(D,digits=NA)
```

## Arguments

| | |
|---|---|
| D | A Dataset object |
| digits | Optionally truncate the digits of the quantiles, if digits>0 |

## Details

The phenotype D$y and each column of the genotype matrix D$genos are replaced by a permutation of the standard Normal quantiles, to improve the security of the encryption. If digits>0 then in addition only the first few decimal digits of each quantil are kept. If digits=NA then no truncation is performed.

This function should be used after encryption by encrypt.D to add an additional level of protection to a Dataset.

## Value

A Dataset object with quantile-normalised phenotype and genotypes. Other elements of the input Dataset are copied verbatim.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020

## See Also

build.D encrypt.D

`rustiefel` *rustiefel*

## Description

Simulate a random orthogonal matrix of dimensions $m * R$ using the Steiefel manifold

## Usage

```
rustiefel(m, R=m)
```

## Arguments

| | |
|---|---|
| `m` | the number of rows of the simulated matrix |
| `R` | thee number of columns |

## Details

Function adapted from R package rsteifel.

## Note

Function adapted from R package rsteifel https://cran.r-project.org/web/packages/rstiefel/index.html.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153 R package rstiefel https://cran.r-project.org/web/packages/rstiefel/index.html

`safe.scale` *Scale genotypes in a dosage matrix by subtracting the column means and dividing by the column standard deviations*

## Description

Scale genotypes safely taking into account the possibility a genotype may have zero variance.

## Usage

```
safe.scale(mat)
```

## Arguments

| | |
|---|---|
| `mat` | A numeric matrix with no missing values |

## Details

Scales the columns of mat by substracting the column mean and dividing by the column standard deviation. If the standard deviation is zero the column is set to zero.

## Value

A matrix with the same dimensions as mat in which each column has been scaled.

## Author(s)

Richard Mott

## References

Mott et al Genetics 2020 https://doi.org/10.1534/genetics.120.303153

# Index