Package 'SoyNAM'

January 21, 2021

Type Package

Title Soybean Nested Association Mapping Dataset

Version 1.6.1

Date 2021-01-21

Author Alencar Xavier, William Beavis, James Specht, Brian Diers, Rouf Mian, Reka Howard, George Graef, Randall Nelson, William Schapaugh, Dechun Wang, Grover Shannon, Leah McHale,
Perry Cregan, Qijian Song, Miguel Lopez, William Muir, Katy Rainey.

Maintainer Alencar Xavier <alenxav@gmail.com>

Description Genomic and multi-environmental soybean data. Soybean Nested Association Mapping (SoyNAM) project dataset funded by the United Soybean Board (USB). BLUP function formats data for genome-wide prediction and association analysis.

License GPL-3

Imports NAM, lme4, reshape2

Depends R (>= 3.2)

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-01-21 15:40:02 UTC

R topics documented:

SoyNAM-package *Soybean Nested Association Mapping Dataset*

Description

Genomic and multi-environmental soybean data. Soybean Nested Association Mapping (SoyNAM) project dataset funded by the United Soybean Board (USB). BLUP function formats data for genome-wide prediction and association analysis.

Details

Package:	NAM
Type:	Package
Version:	1.6.1
Date:	2021-01-20
License:	GPL-3

Author(s)

Alencar Xavier, William Beavis, James Specht, Brian Diers, Rouf Mian, Reka Howard, George Graef, Randall Nelson, William Schapaugh, Dechun Wang, Grover Shannon, Leah McHale, Perry Cregan, Qijian Song, Miguel Lopez, William Muir, Katy Rainey.

Maintainer: Alencar Xavier <alenxav@gmail.com>

See Also

Functions: BLUP, data(soybase), data(soynam) and data(soyin).

BLUP *Best Linear Unbias Predictor*

Description

Genetic values for a given trait computed by REML.

Usage

```
BLUP(trait="yield",family="all",env="all",
     MAF=0.05,use.check=TRUE,impute="FM",rm.rep=TRUE)
```

Arguments

trait	Character. Trait of interest. The options are: "yield" (grain yield in Kg/ha), "maturity" (days to maturity), "height" (plant height in cm), "lodging" (lodging score from 1 to 5), "protein" (protein percentage in the grain), "oil" (oil percentage in the grain), "size" (seed size = mass of 100 seeds in grams) and "fiber" (fiber percentage in the grain).
family	Numberic vector or "all". Which SoyNAM families to use.
env	Numberic vector or "all". Which environments to use. The environments are coded as follows: 1 (IA_2012), 2 (IA_2013), 3 (IL_2011), 4 (IL_2012), 5 (IL_2013), 6 (IN_2012), 7 (IN_2013), 8 (KS_2012), 9 (KS_2013), 10 (MI_2012), 11 (MO_2012), 12 (MO_2013), 13 (NE_2011), 14 (NE_2012), 15 (OHmc_2012), 16 (OHmc_2013), 17 (OHmi_2012) and 18 (OHmi_2013).
MAF	Numeric. Minor allele frequency threshold for the markers.
use.check	Logical. If TRUE, it includes a control term as fixed effect in the model.
impute	NULL, 'RF' of 'FM'. If 'RF', it imputes missing genotypes using random forest. If 'FM' is imputes missing genotypes using a forward Markov Chain algorithm, filling missing loci with the most likely genotype based on the previous marker.
rm.rep	Logical. If TRUE, it removes replicated genotypes. Genotypes are treated as identical when the genotypes are more than 95 percent identical. This argument requires imputed genotypes.

Details

This function uses the raw dataset ($data(soynam)$), allowing user-defined data quality control for genotypes and BLUPs of genetic values.

The algorithm start from selecting the chosen families and environment that will be used for the best linear unbias predictor (BLUP). The BLUP values are calculates based on the following model:

$Trait = Control + Environment + Genotype$)

Where control is a covariate set as fixed effect based on the checks of each set (microenvironment); Environment is a random effect that represents the combination of location and year; and Genotype is the random effect associated to the lines. The BLUP values are the regression coefficients corresponding to the Genotype effect. The BLUP is calculated using the R package lme4 (Bates 2010) using REML.

If checks are used as covariate (use.check=TRUE), then the best linear unbias estimator (BLUE) of the check effects is assigned to each set as a micro-environmental control. Each set had between one and five controls, including the SoyNAM parents and five other cultivars. These genotypes are normalized by environment and the BLUE of each set is calculated. All genotypes in a same set will have the same check effect.

Value

This function returns a list with four objects. A numeric vector with the BLUP solution of the phenotyes ("Phen"); the corresponding genotypes ("Gen"); a vector with the respective family ("Fam"); and a numeric vector with the number of SNPs per cromosome ("Chrom"). The output of this fuction has the exact input format for the NAM package (Xavier et al. 2015) to perform genome-wide association analysis.

Author(s)

Alencar Xavier

References

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. URL http://lme4.r-forge.r-project.org/book.

Xavier, A., Xu, S., Muir, W. M., & Rainey, K. M. (2015). NAM: association studies in multiple populations. Bioinformatics, 31(23), 3862-3864.

Examples

```
Test=BLUP(trait="yield",family=2:3,env=1:2)
```

`Dataset` *Datasets*

Description

Genotypes and phenotypes from quality assured (soybase) or original (soynam) datasets. An additional dataset containing yield components (soyin) collected at Purdue University is also available. See the section "details" for the description of data objects.

The SoyNAM population is a nested association mapping panel that comprises more than 5000 recombinant inbred lines (RILs), including determinate, indeterminate, and semi-determinate genotypes from maturity groups (MG) ranging from late MG II to early MG IV, derived from 40 biparental populations, where progenies were not exposed to selection. Each biparental population approximately contains 140 individuals and all families share the cultivar IA3023 as the standard parent. From the other 40 founder parents, 17 lines are elite public germplasm from different regions, 15 have diverse ancestry and 8 are plant introductions. The SoyNAM population was designed to dissect the genetic architecture of complex traits and to map yield-associated quantitative trait loci (QTL) using a diverse panel.

Parental lines were sequenced to derive the SNP allele calls. A total of 5303 SNP loci were selected with the criterion of maximizing the number of families segregating for those loci. The SNPs were used to build the SoyNAM 6K BreadChip SNP array using the Illumina Infinium HD Assay platform (Song et al. 2017). Among those SNPs, a subset of 4312 markers were selected by the SoyNAM group as quality-assured based on proportion of missing loci and correct segregation patterns. Both raw and quality assured genotypes are available in the R package SoyNAM.

Usage

```
data(soybase)
data(soynam)
data(soyin)
```

Details

Datasets of the SoyNAM project, original and quality assured (QA) versions. Data was downloaded on November 16th 2015 from the soynam website. The data collected in Indiana was collected in 2013-2014 and made available on January 2018. Studies performed on the entire dataset with additional detail about the experimental settings include Diers et al. (2018) and Xavier et al. (2018).

Genotypic matrices are named "gen.raw" and "gen.qa" for the raw and QA versions, respectively. In each dataset, phenotypes are allocated into two objects, one with the lines ("data.line") and one with checks and parents ("data.checks"). Information on data objects include year, location, environment (combination of year and location), strain, family, set (set in each environment), spot (combination of set and environment), height (in centimeters), R8 (number of days to maturity), planting date (501 represents may 1), flowering (701 represents july 1), maturity (901 represents september 1), lodging (score from 1 to 5), yield (in Kg/ha), moisture, protein (percentage in the seed), oil (percentage in the seed), fiber (percentage in the seed), seed size (in grams of 100 seeds).

The dataset including yield components collected at Purdue University (West Lafayette, Indiana) was used to investigate genomic prediction (Xavier et al. 2016) and interaction among traits (Xavier et al. 2017). This dataset contains the genotypic information in the matrix "gen.in", with missing values imputed using the software MaCH (Li et al. 2010). Similar to the datasets previously described, phenotypes are allocated into two objects, lines ("data.line.in") and checks ("data.checks.in"). Information on these data objects include year, location, environment (combination of year and location), strain, family, set (set in each environment), spot (combination of set and environment), the spatial coordinates of the field plots (BLOCK, ROW and COLUMN), plant height (in centimeters), R1 (number of days to flowering), R8 (number of days to maturity), lodging (score from 1 to 5), yield (in bu/ac), leaf shape (ratio length:width), number of nodes in the main stem, number of pods in the main stem, number of pods per node, average canopy coverage, rate of canopy coverage, growing degree day to flowering (GDD_R1), growing degree day to maturity (GDD_R1), and length of reproductive period in terms of growing degree day (GDD_REP).

Author(s)

Alencar Xavier

References

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., ... & Shannon, G. (2018). Genetic Architecture of Soybean Yield and Agronomic Traits. G3: Genes, Genomes, Genetics, g3-200332.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology, 34(8), 816-834.

Song, Q., Yan, L., Quigley, C., Jordan, B.D., Fickus, E., Schroeder, S., Song, B., An, Y. Q. C., Hyten, D., Nelson, R., Rainey, K. M., Beavis, W. D., Specht, J. E., Diers, B. W., Cregan, P. (2017). Development and Genetic Characterization of the Soybean Nested Association Mapping (NAM) Population. Plant Genome. 10(2):1-14.

Xavier, A., Muir, W. M., & Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes, Genetics, 6(8), 2611-2616.

Xavier, A., Hall, B., Casteel, S., Muir, W., & Rainey, K. M. (2017). Using unsupervised learning techniques to assess interactions among complex traits in soybeans. Euphytica, 213(8), 200.

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., ... & Nelson, R. (2017). Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes, Genetics, g3-300300.

Examples

```
data(soynam)
data(soybase)
data(soyin)
```

Index