Usage of corvar.R

corvar.R is an R script which conducts analysis of ‘corvar’, i.e., SNP correlation with variance of local gEBV: and as described in the manuscript ‘Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations’. was the amount of local gEBV variance explained by each variant, was the local gEBV variance, was the squared correlation between the vector of local gEBV () and the vector of the genotype allele count of the variant (). K is the number of traits analysed.

Usage:

Rscript --vanilla corvar.R <../yourpath/filename\_of\_genotype.gz> <../yourpath/filename\_of\_local\_gebv.gz> <../yourpath/filename\_of\_segment\_information.gz> <../yourpath\_of\_output>

Usage with the test data inputs (can be downloaded at <https://figshare.com/s/9e643b192bc816d75a5e>):

Rscript --vanilla corvar.R test.geno.gz test.lgebv.gz test.lgseg.gz test

Test datasets show the format of the input files:

test.geno.gz:

ID V2 V3 V4 V5

A1000 0 1 2 1

A1001 0 1 2 1

A1002 0 0 2 0

A1003 0 1 2 1

test.geno.gz is the 1st argument of the Rscript. This is a genotype count file where the 1st column is animal ID and other columns are genotype counts of 0, 1, 2 for each SNP (each column). These counts are based on the alternative allele of the SNP.

test.lgebv.gz:

seg ID tr01 tr02 tr03

1 A1000 -0.003144169 0.000529620233333333 -0.0004167362

2 A1000 -0.000253421933333333 0.0001216226 7.311196e-05

4 A1000 0.001782575 0.00254010366666667 0.003247314

5 A1000 -0.00144530023333333 -0.0019791532 0.000135896263333333

test.lgebv.gz is the 2nd argument of the Rscript. The 1st column is the order (name) of segments on that chromosome. SNPs within these segments are used to calculate the ‘local genomic estimated breeding value’ or local gEBV. The 2nd column is the animal ID (corresponding to the ID column in the gest.geno.gz). From the 3rd column, these values are the local gEBV per individual per segment for each trait (each column represents data of one trait). So there will be No. of individuals × No. of segments lines of rows in this file. The local gEBV was calculated using the conventional gEBV methods for each trait (e.g., 1), except that the variants used to calculate the local gEBV were from each segment (e.g., 50kb): , where was the local gEBV of the segment *l*, was the design matrix of marker genotypes for variant 1 to variant *n* within the segment *l*, and was the variant effects from the training dataset. See details in the manuscript.

test.lgseg.gz:

SNP order chr bp seg

Chr25:179955 1 25 179955 1

Chr25:181719 2 25 181719 1

Chr25:183050 3 25 183050 1

Chr25:190480 4 25 190480 1

test.lgseg.gz is the 3rd argument for the Rscript. The first column is the ID of SNPs, the 2nd column is the order of SNPs, which corresponds to the column name minus 1 in the ‘test.geno.gz’; the 3rd column is the chromosome number of the SNP, 4th column is the position of the SNP and the 5th column is the segment with which the SNP belong to; the segment number corresponds to the ‘seg’ column in test.lgebv.gz.

if you run Rscript --vanilla corvar.R test.geno.gz test.lgebv.gz test.lgseg.gz test; the script produces an output of ‘test.segcorvar.txt.gz’ (‘test’ is the prefix of output you specified to the Rscript, ie, the 4th argument):

order corvar SNP chr bp seg

1 0 Chr25:179955 25 179955 1

2 1.18308534534575e-08 Chr25:181719 25 181719 1

3 3.29175483369768e-08 Chr25:183050 25 183050 1

4 1.06149887020587e-08 Chr25:190480 25 190480 1

In the output, the 1st column is the order of SNPs, the 2nd column is the metric of ‘corvar’, SNP correlation with variance of local gEBV, based on all traits included in test.lgebv.gz for each SNP. This metric quantifies the importance of the SNP based on its correlation with the local gEBV variance from all traits analysed. Column 3-5 are the SNP information, and the 6th column is the segment with which the SNP belongs to.

Please note that if you have a large number of individuals and variants, it is strongly recommended that you do this analysis chromosome by chromosome, or parts by parts of the genome.

Reference:

1. Kemper, K.E. *et al.* Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution* **47**, 29 (2015).