Usage of segsnpClusterRank.R

segsnpClusterRank.R is an R script that applies to a set of SNPs from a defined chromosome segment, e.g., 1Mb. It takes files of 1] multi-trait GWAS effects of this set of SNPs, 2] the overall ranking of this set of SNPs (e.g., functional importance such as the FAETH score 1) and 3] the LD table of this set of SNPs (calculated using plink: <https://www.cog-genomics.org/plink/1.9/ld#ld_flag>) to conduct cluster analysis of this set of SNPs based on the value of .

where was the correlation across 34 traits between the t values (beta/se from GWAS described above) of and ; was the LD assessed by the correlation between the genotypes of and . The formula is: , where was the correlation across multiple traits between the t values (beta/se from GWAS described above) of and ; was the LD assessed by the correlation between the genotypes of and (See manuscript ‘Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations’). After the clusters are formed between this set of SNPs for that chromosome segment, top variants are selected within each determined cluster given the ranking of the functional importance.

NB: the use of segsnpClusterRank.R is for a set of SNPs from one chromosome segment. Therefore, to complete analysis for a chromosome, this segsnpClusterRank.R shall be repeatedly used across all segments of the chromosome. Each chromosome segment should have some overlap, ideally with half of the size of the segment. For example, if your segment size is 1Mb, then the size of overlap should be 0.5Mb. Partitions of chromosome can be easily done with plink1.9 by giving it different SNP coordinates. Then you can calculate an LD table based on this set of SNPs (\*.ld) which will be used as an input of the Rscript (some instructions detailed in below). Because you could have many SNPs within each chromosome and/or segment, for a whole genome or chromosome analysis, it is strongly recommended to do the analysis on HPC and split the jobs into segments/chromosomes. Pruning and/or clumping with plink is also recommended if you have too many SNPs in very high LD.

Usage:

Rscript --vanilla segsnpClusterRank.R <../yourpath/filename\_of\_multi\_trait\_snp\_effect.gz> <../yourpath/filename\_ranking\_of\_snps.gz> <../yourpath/filename\_of\_plink\_ld\_file\_of\_segment\_snps.ld> <ld\_cutoff\_for\_clustering> <proportion\_of\_top\_snps\_to\_keep> <../yourpath/prefix\_of\_output> <Number\_of\_cores\_for\_analysis>

Usage with the test data inputs (can be downloaded from <https://figshare.com/s/f99ca06e46fc4c2c911b>):

Rscript --vanilla segsnpClusterRank.R test.snpeff.txt.gz test.snprank.txt.gz test.ld 0.95 0.5 test 2

Test datasets show the format of the input files:

test.snpeff.txt.gz:

SNP tr01 tr02 tr03 tr04

Chr30:15045472 -0.181514 -1.749 -0.467912 2.12553

Chr30:15067241 -0.683267 -2.84746 2.88303 2.43342

Chr30:15081372 0.688864 -1.14447 0.178485 1.30934

Chr30:15090245 -0.164506 -2.06394 2.97858 1.85093

test.snpeff.txt.gz is the 1st argument of the Rscript and contains the multi-trait SNP effect (t value, beta/se) from GWAS. The 1st column is the SNP ID and the rest columns are the SNP t value for each trait.

test.snprank.txt.gz:

SNP snpRank

Chr30:16175247 1824

Chr30:16062873 1831.5

Chr30:16179688 1831.5

Chr30:16190543 1831.5

test.snprank.txt.gz is the 2nd argument of the Rscript and contains the functional importance ranking of SNPs. This ranking is based on all SNPs or SNPs from the targeted chromosome. Note that this ranking is some kind of functional importance of the SNPs that should not be highly correlated with their GWAS effects. An example of such ranking is the Functional-And-Evolutionary Trait Heritability (FAETH) ranking 1 (used in the manuscript Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations).

test.ld:

CHR\_A BP\_A SNP\_A CHR\_B BP\_B SNP\_B R

30 15045472 Chr30:15045472 30 15067241 Chr30:15067241 0.0575144

30 15045472 Chr30:15045472 30 15081372 Chr30:15081372 0.727995

30 15045472 Chr30:15045472 30 15090245 Chr30:15090245 0.0484118

30 15045472 Chr30:15045472 30 15095356 Chr30:15095356 -0.0048174

test.ld is the 3rd argument of the Rscript and is the output from plink ‘--r' for a given set of SNPs. This can be obtained by the following function as an example: ./plink --bfile test --keep-allele-order --r --extract seg1.snp.list --out seg1.snp --threads 2; where ‘test’ is the prefix of the plink binary file (potentially all chromosome/genome SNPs are in there), ‘--r’ is the function to calculate raw inter-variant allele count correlations, ‘--extract seg1.snp.list’ is to specify a list of SNPs from the targeted segment for the calculation; ‘--out’ specifies the output prefix of the LD correlation results of ‘\*.ld’.

The 4th argument of the Rscript is the cutoff value of LD r2. Variants with LD r2 above this value (we used 0.95) are in very high LD and they might bias the clustering. Thus they were ranked and selected directly based on their functional importance. Variants with LD r2 below this value were included in the clustering analysis.

The 5th argument of the Rscript is the proportion of top variants after the clustering. We have used 0.5 and this means that the variants ranked within top 50% based on their functional importance are selected within each cluster. If this value is 0.1 then the variants ranked within top 10% are selected.

The 6th argument is the prefix of output (‘test’ a in the example).

The 7th argument specifies the number of cores can be used. This is to speed-up the calculation of large correlation matrix.

The Rscript will generate 3 outputs (with prefix of ‘test’ as specified by the user):

test.selectsnp.txt.gz:

selected snpRank

Chr30:15105508 245765

Chr30:15450895 34775

Chr30:15462954 50945.5

Chr30:15544886 60261.5

The 1st column of test.selectsnp.txt.gz is the ID of top SNPs selected by the clustering and ranking method. The 2nd column is their global functional ranking of SNPs (from .test.snprank.txt.gz)

test.perfldsnp.txt.gz:

selected Var1.rank Var2 Var2.rank ld p.cor pr

Chr30:15462954 50945.5 Chr30:15488812 50945.5 1 1 1

Chr30:15544886 60261.5 Chr30:15553321 60261.5 1 1 1

test.perfldsnp.txt.gz contains the pairwise SNPs which have perfect LD and thus can not be properly differentiated by the clustering method. Column ‘ld’ is the LD correlation between the two SNPs, column ‘p.cor’ is the correlation of SNP effect on multiple traits (10 traits in this case) between the 2 SNPs and column ‘pr’ is the metric of between the two SNPs. As shown in the table, all correlations are 1 so the we cant differentiate the 2 SNPs given the data we have. For SNP selection purpose we have kept them to see if other analysis can distinguish them.

test.clustsnp.txt.gz:

SNP k snpRank rk

Chr30:15045472 1 1457455 0.857142857142857

Chr30:15067241 1 1643763 1

Chr30:15081372 1 682690.5 0.357142857142857

Chr30:15090245 1 1052317 0.642857142857143

test.clustsnp.txt.gz contains detailed clustering results for variants that participated in the clustering and ranking. The 1st column is the SNP ID, 2nd column is the name of cluster that they were assigned, the 3rd is their global functional importance ranking (from test.snprank.txt.gz) and the 4th column is the within cluster percentile ranking of SNPs. So if you set 0.5 as the 5th argument of the Rscript, it would selected variants with the percentile ranking (‘rk’ column) <= 0.5.

Reference:

1. Xiang, R. *et al.* Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* **116**, 19398-19408 (2019).