MARV

INTRODUCTION

MARV (Multi-phenotype Analysis of Rare Variants) is an implementation of a method described in Kaakinen *et al.*[1] to perform multi-phenotype analysis of rare-variants. The method is based on joint analysis of multiple phenotypes and accumulation of minor alleles of rare or low-frequency markers discovered through dense genotyping, resequencing data or imputation. Association analyses are based on gene- or other pre-defined regions, determined by analyst.

[1] Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin M-R, Morris AP, Prokopenko I. A rare variant test for high-dimensional data. *European Journal of Human Genetics* 2017;25(8):988-994.

CITATION

Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin M-R, Morris AP, Prokopenko I. A rare variant test for high-dimensional data. *European Journal of Human Genetics* 2017;25(8):988-994.

Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin M-R, Morris AP, Prokopenko I. MARV: A Tool for Genome-Wide Multi-phenotype Analysis of Rare Variants. *BMC Bioinformatics* 2017; 18:110.

DOWNLOAD

The following files are provided:
MARV_v1.0.4.zip
ucsc_genes_b37.txt.zip (Gene list, positions dbsnp37)
coding_markers_b37.list.zip (Coding region marker extraction list, positions dbsnp37)
nonsynonymous_markers_b37.list.zip (Non-synonymous marker extraction list, positions dbsnp37)

INSTALLATION

Copy MARV_v*.zip file into your computer and unzip the file:
unzip MARV_v*.zip

To compile MARV program, type the following command in the folder where files have been unpacked:
make

The program can be run by typing:
./MARV

INPUT FILES

For running MARV, you need input files in SNPTESTv.2 format and a GENELIST file. SNPTEST file formats are described here.

GENOTYPE file:
1 rs1 11 A T 1 0 0 1 0 0 1 0 0
1 rs2 210 A T 0 1 0 1 0 0 1 0 0
1 rs3 300 A T 1 0 0 1 0 0 1 0 0
(Genotype file can be gzipped, if it has *.gz extension)

SAMPLE file:
Sample_id Subject_id Missing Gender Phenotype Phenotype
Q 0 0 0 D B P
1 1 0 1 1 4.1
2 2 0 1 1 4.2
3 3 0 1 0 4.3

This file contains one covariate (Gender) and two phenotypes: first one is binary and second one is a continuous phenotype.

GENELIST file:
A1 1 11 111
A5 1 2500 27300
A13 1 14 3000

Genelist file contains four columns: 1. GENE/region ID 2. chromosome 3. start position in bp 4. end position in bp.

NB! Make sure that the positions in genelist file are from the same dbSNP version as in the GEN file.

RUNNING MARV

Command line options:

```
./MARV  [--debug] [--print_covariance] [--print_complex]
[--betas] [--print_all] [--remove_missing] --pheno_name <string>
[-f <double>] [-r <double>] [--call_thresh <double>]
[--imp_thresh <double>][--missing_code <double>]
[--extract_markers <string>][--extract_samples <string>]
[--exclude_markers <string>][--exclude_samples <string>]
-m <threshold|expected> -x <string>
[-o <string>] [--chr <string>] -g <string> -s <string>
[--version] [-h]
```

Where:

--debug
Debug mode enabled

--print_covariance
  Print covariance matrix data for the model with all phenotypes
  (default OFF)

--print_complex
  Print only the model with all phenotypes (default OFF)

--betas
  Print effect size and stderr info (default OFF)

--print_all
  Print results for all models (default OFF)

--remove_missing
  Remove sample if any of the phenotype values is missing (default OFF)

--pheno_name <string> (accepted multiple times)
  (required) Name of phenotype to use (use this command multiple times
  i.e. --pheno_name BMI --pheno_name HEIGHT etc.)

-f <double>, --flanking <double>
  This specifies flanking region size in kb (default 0 kb)

-r <double>, --rare_thresh <double>
  Minor allele cut-off for defining rare variants (default 0.05)

--call_thresh <double>
  Call-rate threshold for the best guess genotypes (default 0.9)

--imp_thresh <double>
  Imputation score threshold for including markers (default 0.4)

--missing_code <double>
  This specifies the coding for missing data (default NA)

--extract_markers <string>
  This specifies file with extracted markers

--extract_samples <string>
  This specifies file with extracted samples

--exclude_markers <string>
  This specifies file with excluded markers

--exclude_samples <string>
  This specifies file with excluded samples

-m <threshold|expected>, --method <threshold|expected>
  (required) This option controls how the genotype uncertainty is taken
  into account

-x <string>, --genmap <string>
  (required) This specifies gene map file

-o <string>, --out <string>
  This specifies output files root

--chr <string>
  All markers in genotype file are forced to be from this chromosome.
  Ignoring the first column in genotype file

-g <string>, --gen <string>
  (required) This specifies genotype file

-s <string>, --sample <string>
  (required) This specifies sample file

--, --ignore_rest
  Ignores the rest of the labeled arguments following this flag.

--version
  Displays version information and exits.

-h, --help
  Displays usage information and exits.

OUTPUT FILES

MARV print outs by default the following files: .err file, .log file and .result file. By request a .betas file can also be printed.

.err file is empty if the analysis was successful; otherwise details of the error occurred are printed in the file.

.log file provides details of the data analysed as well as of the markers included in each genomic region, together with their minor allele frequencies.

.results file contains the following columns:

Column name	Description
Gene	Gene ID
RareCount	Number of rare markers in gene region
N	Number of samples in analysis
RareVariantSum	Count of rare alleles found in individuals
TotalMAF	Sum of MAF of all used markers in given gene region
AverageMAF	Average MAF of used markers in given gene region (total_maf / marker_count)
PhenotypeCount	Number of phenotypes used for the modelling
Mask	Which of the phenotypes were used (1) and which not (0) out of the all phenotypes given
LogLikelihood	Log likelihood of the model
nullLogLikelihood	Log likelihood of the null model
LikelihoodRatio	Likelihoodratio
Pvalue	P-value of the model, uncorrected for any multiple testing
BIC	Bayesian information criterion
BICnull	Bayesian information criterion for the null model
Model	Included phenotypes
sortedModel	Included phenotypes sorted

If print_covariance option is used, the following columns are also output:

Column name	Description
beta_k	Effect size for phenotype k, k=1,…n
se_k	Standard error of the effect size for phenotype k, k=1,…,n
cov_j_k	The inverted covariances between phenotypes j and k, j,k=1,..,n

.betas file contains the following columns:

Column name	Description
Gene	Gene ID
RareCount	Number of rare markers in gene region
N	Number of samples in analysis
RareVariantSum	Count of rare alleles found in individuals
TotalMAF	Sum of MAF of all used markers in given gene region
AverageMAF	Average MAF of used markers in given gene region (total_maf / marker_count)
Model	Included phenotypes
Model_member	Phenotype for which the beta and se are reported
beta	Effect size
se	Standard error of the effect size