BMC Bioinformatics

SOFTWARE Open Access

MARV: a tool for genome-wide multi-phenotype analysis of rare variants

Marika Kaakinen[1*], Reedik Mägi[2], Krista Fischer[2], Jani Heikkinen[1,3], Marjo-Riitta Järvelin[4,5,6,7], Andrew P. Morris[8] and Inga Prokopenko[1*]

Abstract

Background: Genome-wide association studies have enabled identification of thousands of loci for hundreds of traits. Yet, for most human traits a substantial part of the estimated heritability is unexplained. This and recent advances in technology to produce high-dimensional data cost-effectively have led to method development beyond standard common variant analysis, including single-phenotype rare variant and multi-phenotype common variant analysis, with the latter increasing power for locus discovery and providing suggestions of pleiotropic effects. However, there are currently no optimal methods and tools for the combined analysis of rare variants and multiple phenotypes.

Results: We propose a user-friendly software tool MARV for Multi-phenotype Analysis of Rare Variants. The tool is based on a method that collapses rare variants within a genomic region and models the proportion of minor alleles in the rare variants on a linear combination of multiple phenotypes. MARV provides analyses of all phenotype combinations within one run and calculates the Bayesian Information Criterion to facilitate model selection. The running time increases with the size of the genetic data while the number of phenotypes to analyse has little effect both on running time and required memory. We illustrate the use of MARV with analysis of triglycerides (TG), fasting insulin (FI) and waist-to-hip ratio (WHR) in 4,721 individuals from the Northern Finland Birth Cohort 1966. The analysis suggests novel multi-phenotype effects for these metabolic traits at *APOA5* and *ZNF259*, and at *ZNF259* provides stronger support for association ($P_{TG+FI} = 1.8 \times 10^{-9}$) than observed in single phenotype rare variant analyses ($P_{TG} = 6.5 \times 10^{-8}$ and $P_{FI} = 0.27$).

Conclusions: MARV is a computationally efficient, flexible and user-friendly software tool allowing rapid identification of rare variant effects on multiple phenotypes, thus paving the way for novel discoveries and insights into biology of complex traits.

Keywords: Rare variant analysis, Multi-phenotype analysis, High-dimensional data

Background

In the past decade, genetic locus discovery for human traits and diseases has been advanced via genome-wide association studies (GWAS). Recent improvements in technology to produce genotype data in a very cost- and time-effective manner and powerful easy-to-use software tools have played a major role in these advances, facilitating fast analysis of constantly increasing amounts of data. Clearly, the next advances in the field of genomics will be based on large-scale sequencing and other high-dimensional omics data. A key challenge for successful utilisation of these data lies, once again, in the availability of powerful methods and user-friendly software tools, thus enabling researchers to make rapid discoveries [1].

Large-scale sequencing efforts, such as the 1000 Genomes Project [2] or more recently the UK10K Project [3] and the Haplotype Reference Consortium [4], have enabled better characterization of variation in the human genome, especially in the low-frequency and rare variant range. Here, we denote all variants with minor allele frequency, MAF < 5%, by RVs. Imputation based on variant density detected by these projects yields high-quality genotype data even down to 0.01% allele frequency [5]. Large scale sequencing data generation encourages method and software

* Correspondence: m.kaakinen@imperial.ac.uk; i.prokopenko@imperial.ac.uk
[1] Department of Genomics of Common Disease, Imperial College London, London W12 0NN, UK
Full list of author information is available at the end of the article

development for elucidating RV effects, since traditional single-variant methods are underpowered to detect RV associations. Several methods and related software tools have been proposed, including burden tests using collapsing techniques, variance-component tests and combinations of the two [6].

There has also been increasing interest in addressing analysis of high-dimensional phenotypic and omic data, such as metabolomics, in relation to human genome variation. Multi-phenotype analysis (MPA), i.e. joint analysis of multiple phenotypes, is an example of recent developments in the field. Several methods and related software for single-variant MPA, including Bayesian and frequentist approaches, have recently been published [7]. The MPA approach is motivated by several factors: 1) it boosts power for locus discovery [8–11]; 2) it provides more precise parameter estimates [12]; and 3) it has biological advantages including the possibility to identify multi-phenotype effects, including pleiotropy [13], when one locus affects multiple phenotypes. The power improvement by the MPA approach is especially relevant from a computational point of view, because to enable the discovery of further loci for complex traits, the analyses will need to be based on hundreds of thousands of individuals, such as those available from the UK Biobank and other new large-scale efforts based on sequencing. Storage and computational load for such amounts of data will pose a challenge, and alternative strategies for boosting power for locus discovery other than that of increasing sample size, clearly bring an enormous advantage.

We propose a novel tool MARV for RV MPA, which enables joint analysis of both large-scale high-dimensional genomic and phenotypic data. It extends the burden test for RVs to high-dimensional phenotypic data by applying the MPA approach. Recently, methods designed for MPA of RVs have been proposed [14–16], but these have several limitations regarding scalability and ability to combine continuous and discrete phenotypes, and more importantly, the associated software: they either lack an easy user-interface or are computationally inefficient – key features to facilitate fast discoveries. Our software tool MARV enables analysis of both continuous and binary phenotypes, as well as genotyped, imputed or sequenced data. MARV is computationally efficient for large-scale data. From a user point of view, it enables standard formats of data as used in other GWAS software, and the analyses are run using a command line interface, also familiar from widely used GWAS software such as Plink [17] and SNPTEST [18], thus enabling researchers quickly and effortlessly to transit from the standard single variant, single phenotype GWAS to region-based analysis of multiple phenotypes.

Implementation

The method on which MARV is based is briefly introduced in Methods, and is extensively described, including power simulations, elsewhere [19] (Methods). MARV is written in C++ and has a command line user-interface. A single run of MARV consists of just one step and the required input files, commands and the resulting output files are described below.

Data input and commands

MARV requires three files for a successful run: sample, genotype and genomic region input files (Fig. 1). The sample and the genotype input files should be in the SNPTEST v2 [18] format. The genomic region file should contain three columns: the name of the region, and the start and the end positions for the region. It is important that the positions in this file correspond to

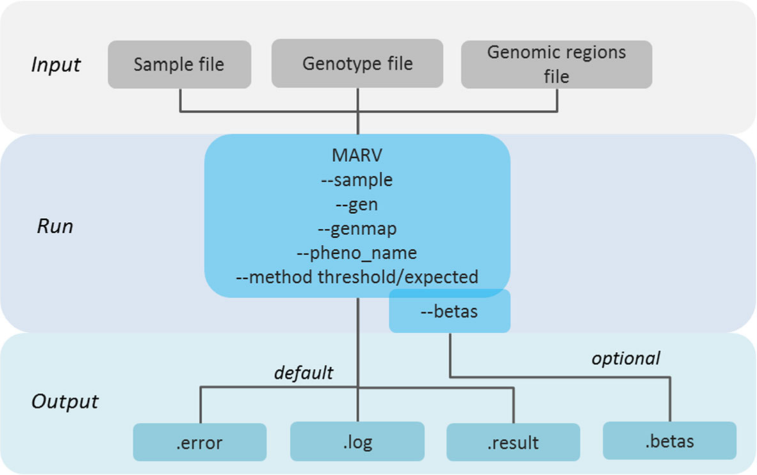


Fig. 1 Workflow of a MARV run including required files, commands and resulting output files

the positions of the genotype file, i.e. the same genome build for these two should be used (Fig. 2).

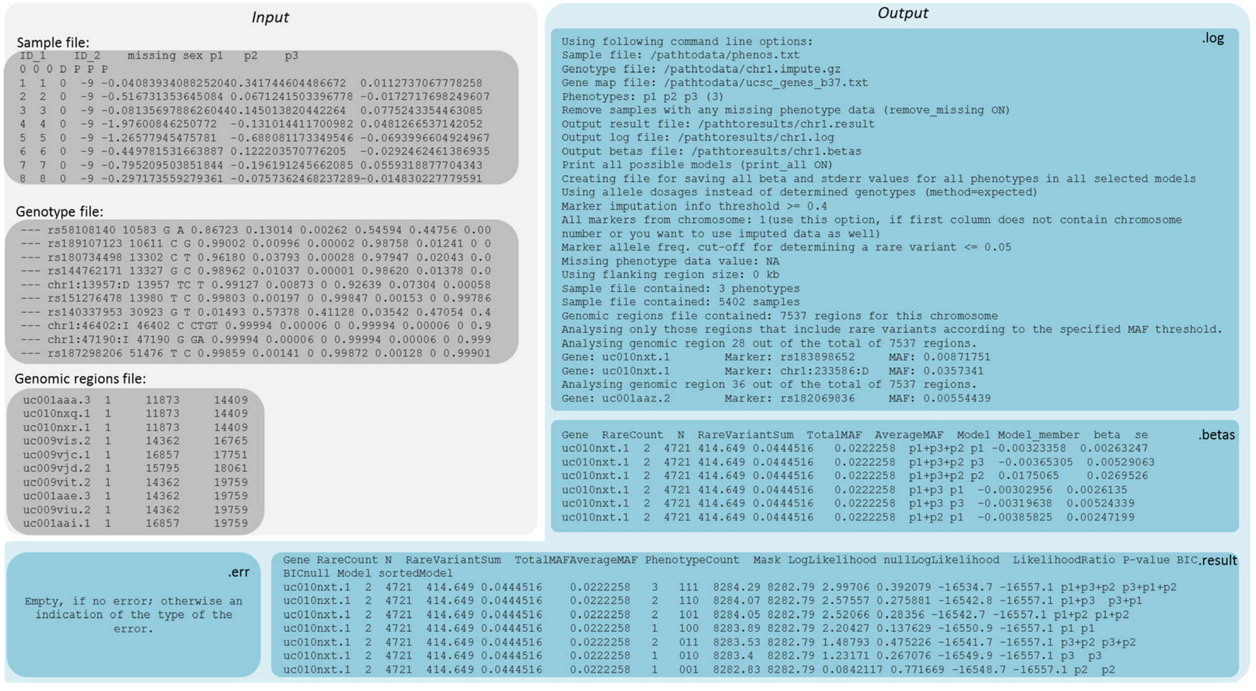
The user then needs to specify the phenotypes to be analysed (−pheno_name), corresponsing to a column name in the sample file, and the method to use for the analysis, i.e. whether to analyse the genotype dosages derived by the software from the imputation probabilities (−method expected) or whether to use the thresholded genotypes based on a pre-defined cut-off (−method threshold, with a cut-off default 0.9 which can be changed with the −call_thres option) (Fig. 1). Additionally, the user may specify several other options, such as individuals or SNPs to extract or exclude from the analysis. It is important to specify the threshold used for the minor allele frequency (−rare_thresh, by default 0.05, which means that variants with minor alleles of frequency < 5% only will be included in the analysis). All the available options of the latest version of MARV can be found from the online manual of MARV.

Data analysis

MARV works across the genome by going through the specified gene regions one by one. Based on the gene boundaries and desired rare variant cut-off, it calculates, for each individual, the proportion of minor alleles at rare variants within the region [20]. After this calculation is performed for all individuals, a linear regression is fitted using the proportion as the outcome and the listed phenotypes as its predictors. The likelihood contribution of each individual is further weighted by the number of successfully genotyped/imputed RVs in the region of interest. For each genomic region, weighted linear regression is performed for all different phenotype combinations, i.e. if a user specifies phenotypes *pheno_a* and *pheno_b*, three different models for the proportion are fitted with the following predictor combinations: 1) *pheno_a* + *pheno_b*, 2) *pheno_a*, 3) *pheno_b*. MARV calculates the Bayesian Information Criterion (BIC) for each model to help the user in identifying the best fitting phenotype combination.

Output files

MARV produces three files by default:.error file,.log file and.result file (Fig. 1). The error file will be empty if the run was completed successfully; otherwise details about problems during the run are reported (Fig. 2). The log file will give specific details of the analysis, including the number of samples in the sample file and genotype file, and the number of phenotypes used for the analysis. It will also include the variants included for the analysis of each genomic region, along with their MAFs. The results file will include one row for results per each genomic region. If the user specifies printing of all the possible model combinations (−print_all) there will be as many rows per gene as there were different model combinations fitted. This file will inform the log likelihood and BIC of the model as well as the *P*-value for each model. We note that the *P*-value is uncorrected for any multiple



Input

Sample file:

```
ID_1    ID_2    missing sex p1    p2    p3
0 0 0 D P P P
1  1  0  -9 -0.04083934088252040.341744604486672  0.0112737067778258
2  2  0  -9 -0.516731353645084 0.0671241503396778 -0.0172717698249607
3  3  0  -9 -0.08135697886260440.145013820442264  0.0775243354463085
4  4  0  -9 -1.97600846250772  -0.131014411700982 0.0481266537142052
5  5  0  -9 -1.26577945475781  -0.688081173349546 -0.0693996604924967
6  6  0  -9 -0.449781531663887 0.122203570776205  -0.0292462461386935
7  7  0  -9 -0.795209503851844 -0.196191245662085 0.0559318877704343
8  8  0  -9 -0.297173559279361 -0.0757362468237289-0.01483022777959l
```

Genotype file:

```
--- rs58108140 10583 G A 0.86723 0.13014 0.00262 0.54594 0.44756 0.00
--- rs189107123 10611 C G 0.99002 0.00996 0.00002 0.98758 0.01241 0 0
--- rs180734498 13302 C T 0.96180 0.03793 0.00028 0.97947 0.02043 0.0
--- rs144762171 13327 G C 0.98962 0.01037 0.00001 0.98620 0.01378 0.0
--- chr1:13957:D 13957 TC T 0.99127 0.00873 0 0.92639 0.07304 0.00058
--- rs151276478 13980 T C 0.99803 0.00197 0 0.99847 0.00153 0 0.99786
--- rs140337953 30923 G T 0.01493 0.57378 0.41128 0.03542 0.47054 0.4
--- chr1:46402:I 46402 C CTGT 0.99994 0.00006 0 0.99994 0.00006 0 0.9
--- chr1:47190:I 47190 G GA 0.99994 0.00006 0 0.99994 0.00006 0 0.999
--- rs187298206 51476 T C 0.99859 0.00141 0 0.99872 0.00128 0 0.99901
```

Genomic regions file:

```
uc001aaa.3  1      11873      14409
uc010nxq.1  1      11873      14409
uc010nxr.1  1      11873      14409
uc009vis.2  1      14362      16765
uc009vjc.1  1      16857      17751
uc009vjd.2  1      15795      18061
uc009vit.2  1      14362      19759
uc001aae.3  1      14362      19759
uc009viu.2  1      14362      19759
uc001aai.1  1      16857      19759
```

Output

.log

```
Using following command line options:
Sample file: /pathtodata/phenos.txt
Genotype file: /pathtodata/chr1.impute.gz
Gene map file: /pathtodata/ucsc_genes_b37.txt
Phenotypes: p1 p2 p3 (3)
Remove samples with any missing phenotype data (remove_missing ON)
Output result file: /pathtoresults/chr1.result
Output log file: /pathtoresults/chr1.log
Output betas file: /pathtoresults/chr1.betas
Print all possible models (print_all ON)
Creating file for saving all beta and stderr values for all phenotypes in all selected models
Using allele dosages instead of determined genotypes (method=expected)
Marker imputation info threshold >= 0.4
All markers from chromosome: 1(use this option, if first column does not contain chromosome
number or you want to use imputed data as well)
Marker allele freq. cut-off for determining a rare variant <= 0.05
Missing phenotype data value: NA
Using flanking region size: 0 kb
Sample file contained: 3 phenotypes
Sample file contained: 5402 samples
Genomic regions file contained: 7537 regions for this chromosome
Analysing only those regions that include rare variants according to the specified MAF threshold.
Analysing genomic region 28 out of the total of 7537 regions.
Gene: uc010nxt.1        Marker: rs183898652     MAF: 0.00871751
Gene: uc010nxt.1        Marker: chr1:233586:D   MAF: 0.0357341
Analysing genomic region 36 out of the total of 7537 regions.
Gene: uc001aaz.2        Marker: rs182069836     MAF: 0.00554439
```

.betas

```
Gene  RareCount  N  RareVariantSum  TotalMAF  AverageMAF  Model Model_member  beta  se
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p3+p2 p1 -0.00323358  0.00263247
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p3+p2 p3  -0.00365305  0.00529063
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p3+p2 p2  0.0175065    0.0269526
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p3 p1  -0.00302956  0.0026135
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p3 p3  -0.00319638  0.00524339
uc010nxt.1  2  4721 414.649 0.0444516   0.0222258  p1+p2 p1  -0.00385825  0.00247199
```

.err

```
Empty, if no error; otherwise an
 indication of the type of the
          error.
```

.result

```
Gene RareCount N  RareVariantSum  TotalMAFAverageMAF PhenotypeCount  Mask LogLikelihood nullLogLikelihood  LikelihoodRatio P-value BIC
BICnull Model sortedModel
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   3   111   8284.29 8282.79 2.99706 0.392079 -16534.7 -16557.1 p1+p3+p2 p3+p1+p2
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   2   110   8284.07 8282.79 2.57557 0.275881 -16542.8 -16557.1 p1+p3  p3+p1
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   2   101   8284.05 8282.79 2.52066 0.28356 -16542.7 -16557.1 p1+p2 p1+p2
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   1   100   8283.89 8282.79 2.20427 0.137629 -16550.9 -16557.1 p1 p1
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   2   011   8283.53 8282.79 1.48793 0.475226 -16541.7 -16557.1 p3+p2 p3+p2
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   1   010   8283.4  8282.79 1.23171 0.267076 -16549.9 -16557.1 p3  p3
uc010nxt.1  2  4721  414.649 0.0444516    0.0222258   1   001   8282.83 8282.79 0.0842117 0.771669 -16548.7 -16557.1 p2  p2
```

Fig. 2 Examples of the required input file formats for MARV and the resulting output files

testing. If the user is interested in the effect estimates and their standard errors for each of the model members, i.e. phenotypes included in the fitted model, a separate.betas file can be requested (–betas) (Figs. 1 and 2). A complete list of the columns in the output file with their meanings is provided in the online tutorial of MARV.

Case study

To illustrate the use of MARV across the genome, we have applied it to data from the Northern Finland Birth Cohort 1966 (NFBC1966), which covers over 96% of all births in the two northernmost provinces of Finland in 1966 ($N = 12{,}068$ live-born children) [21]. We included data from 4,721 cohort members who had participated in the 31 year clinical examination and had genetic data as well as data on triglycerides (TG), fasting insulin (FI) and waist-to-hip ratio (WHR). The Ethics Committees of the University of Oulu and Northern Ostrobothnia Hospital District have approved the study. Individuals used for the analyses have provided written, informed consent.

Motivation for the selection of the traits comes from a common variant single-trait GWAS, which has shown an enrichment of FI associations among SNPs preselected on Metabochip for TG and waist phenotypes [22]. For the selected traits, we applied the following criteria: 1) FG: exclude non-fasting individuals and/or those having type 1 or 2 diabetes mellitus or on diabetes treatment or having fasting blood glucose ≥ 7 mmol/l and/or being pregnant, 2) TG: exclude non-fasting and/or individuals known to be on lipid lowering medication. We modelled each trait on sex, body mass index and the first three principal components derived from the genetic data to control for potential population structure. An inverse normal transformation was further applied to the residuals of WHR and TG to reduce skewness.

DNA was extracted from blood samples drawn after overnight fasting at the 31 year clinical examination. Genotyping was performed with the Illumina HumanCNV370-DUO Analysis BeadChip platform at the Broad Institute, USA, with Beadstudio algorithm being used for genotype calling. Detailed genotyping and sample quality control (QC) of the first set of data have been reported before [23]. Additional samples were genotyped afterwards, resulting in 5,402 subjects and 324,896 SNPs available for analysis. The 1,000 Genomes Project "all ancestries" reference panel (March 2012) was used for imputation, resulting in ~38 M SNPs for analysis.

We analysed the transformed residuals in MARV with the method "threshold" (option -m threshold), i.e. genotypes with probability of 0.95 or higher were considered called, whilst all others were considered missing. The gene list from the University of California Santa Cruz (UCSC, NCBI genome sequence build 37, hg19) [24] was used to define gene regions, and a level of significance of 1.67×10^{-6} was adopted based on a Bonferroni correction for 30,000 genes. We analysed all variants irrespective of their annotation across autosomal chromosomes using the following cut-offs: MAF < 5% and imputation quality > 0.4.

Results and discussion

Case study

The three selected phenotypes, FI, TG and WHR, were modestly correlated with each other ($R_{FI\_TG} = 0.37$, $R_{FI\_WHR} = 0.18$, $R_{TG\_WHR} = 0.19$). The multi-phenotype analysis of the three phenotypes revealed genome-wide significant associations covering two gene regions on chromosome 11: at *APOA5* (apolipoprotein A-V) and at *ZNF259* (the zinc finger protein 259, also known as *ZPR1*) genes (Table 1, Figs. 3 and 4). Besides the full model, MARV also provides parameter estimates and tests of associations for each phenotype combination, including the single phenotype models. Therefore, we were able to compare the results from the joint analysis against traditional single phenotype analyses. Additionally, the BIC provided by MARV for each sub-model served for selection of the phenotype combination providing the best fit. At *APOA5*, the best fitting model according to BIC contained TG only ($P = 2.0 \times 10^{-7}$), while at *ZNF259*, the model with FI and TG provided the lowest BIC and hence support for the best fit ($P = 1.8 \times 10^{-9}$) (Table 1). The model with FI and TG provided a lower *P*-value compared to those obtained from univariate models ($P_{TG} = 6.5 \times 10^{-8}$ and $P_{FI} = 0.27$), suggesting that at least the association with FI would have been missed in univariate analyses. The effects of TG and FI on the rare allele load were in opposite directions: while the increase in TG levels was associated with a greater proportion of minor alleles within *ZNF259*, the opposite was true for FI. This was true also for the univariate models (Table 1). All results files outputted by MARV are available as Additional files 1, 2, 3 and 4.

Common variants at these two identified genes have previously been associated with TG, total cholesterol, high-density lipoprotein, low-density lipoprotein, apolipoprotein A1 and B, coronary heart disease, coronary artery disease, plasma viscosity, Lp-PLA2 activity, prostate cancer, and circulating vitamin E levels [25–33]. A recent large-scale GWAS also reported RV associations at *ZNF259* with triglyceride levels [34]. Our analysis pointed to multi-phenotype effects with TG and FI. A recent study in Japanese individuals showed evidence for associations between variation in *ZNF259* and type 2 diabetes [35], making this locus of interest for further investigation in the pathogenesis of the disease. Interestingly though, in our MPA the effects of TG and FI on the rare allele load at *ZNF259* were in opposite

Table 1 Results for loci reaching genome-wide significance in the multi-phenotype rare variant analysis of NFBC1966 ($N = 4{,}721$). Regression coefficients with their standard errors (SE) are reported, followed by the *P*-value and the Bayesian Information Criterion (BIC) for the analysed model. TG, triglycerides; ln(FI), natural logarithm transformed fasting insulin; WHR, waist-to-hip ratio

	APOA5 (Chr 11: 116,660,086-116,663,136)		*ZNF259* (Chr 11: 116,649,276-116,658,739)	
Model	β (SE)	*P*-value; BIC	β (SE)	*P*-value; BIC
TG + ln(FI) + WHR, full model [a]		3.32×10^{-8}; −19877.3		6.3×10^{-8}; −25069.6
TG	0.011 (0.002)	-	0.007 (0.001)	-
ln(FI)	−0.010 (0.004)	-	−0.008 (0.002)	-
WHR	0.027 (0.019)	-	0.010 (0.011)	-
TG + ln(FI)		2.00×10^{-8}; −19883.5		1.8×10^{-9}; −25077.3
TG	0.011 (0.002)	-	0.007 (0.001)	-
ln(FI)	−0.010 (0.004)	-	−0.007 (0.002)	-
TG + WHR		3.34×10^{-7}; −19877.9		4.1×10^{-7}; −25066.5
TG	0.009 (0.002)	-	0.005 (0.001)	-
WHR	0.020 (0.019)	-	0.005 (0.001)	-
ln(FI) + WHR		0.08; −19853.1		0.10; −25427.6
ln(FI)	−0.003 (0.004)	-	−0.003 (0.002)	-
WHR	0.041 (0.019)	-	0.019 (0.011)	-
Univariate [b]				
TG	0.009 (0.002)	9.15×10^{-8}; −19885.1	0.005 (0.001)	6.5×10^{-8}; −25074.7
ln(FI)	−0.002 (0.003)	0.62; −19856.8	−0.002 (0.002)	0.27; −25046.7
WHR	0.038 (0.019)	0.04; −19860.7	0.016 (0.011)	0.16; −25047.5

[a] For a genome-wide joint analysis, the level of significance is $P < 1.67 \times 10^{-6}$ after Bonferroni correction for 30,000 genes

[b] For univariate analysis, the level of significance is $P < 5.56 \times 10^{-7}$ after Bonferroni correction for 30,000 genes and three phenotypes

directions, contrary to our expectations, since elevated TG levels usually correlate with elevated rather than decreased FI levels.

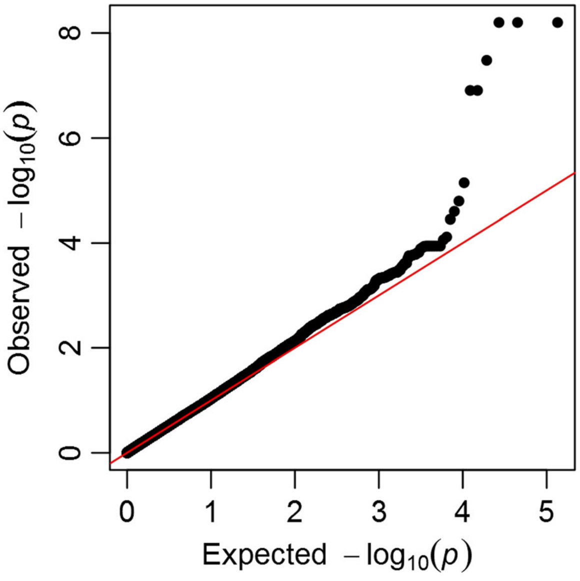


Fig. 3 QQ-plot of MARV analysis results on triglycerides, fasting insulin and waist-to-hip ratio in the NFBC1966

Running time and memory

We measured running time and memory usage of MARV by performing additional analyses on the NFBC1966 data with different number of individuals, phenotypes and on different sized chromosomes. For these analyses, we used 2,405 and 4,809 (i.e. ~double the first) individuals with complete data on eight continuous phenotypes. We analysed a combination of two, four and eight continuous phenotypes and used 1000 Genomes imputed chromosomes 1 and 22 data for the association analyses. All analyses were run and their performance data were collected using Imperial College HPC Cluster. Compute nodes were equipped with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz machine.

The results are summarised in Table 2. We observe that the size of the genomic region to be analysed notably affects the running time. However, there is not a linear relationship between the number of phenotypes and the computation time required. For example, the increase in time for chromosome 1 is just under 3 h (17% from original time) even when the number of

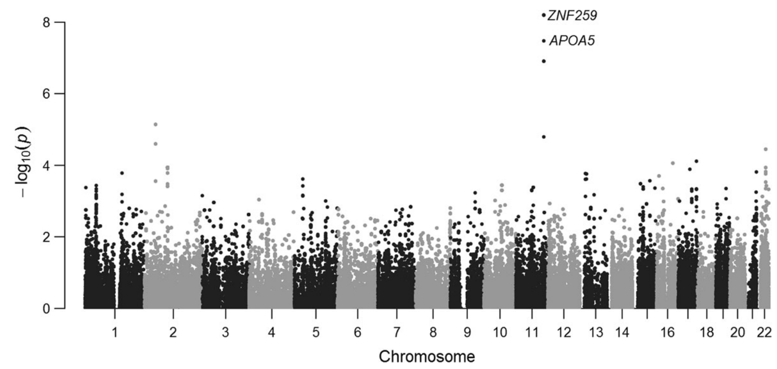


Fig. 4 Manhattan plot of MARV analysis results on triglycerides, fasting insulin and waist-to-hip ratio in the NFBC1966. Genes reaching genome-wide significance ($P < 1.67 \times 10^{-6}$) are annotated

phenotypes is doubled from four to eight and the number of models to be fitted is more than 20-fold. Doubling the sample size roughly triples the runtime.

The memory usage of MARV is more related to the size of the genetic data and number of individuals to analyse rather than the number of phenotypes to analyse. In our example, the peak memory usage was almost constant for all chromosome 1 and 22 analyses when the sample size remained the same, independent of the number of phenotypes in the model (Table 2). Considering the size differences of these two chromosomes (Table 2), we note that the increase in memory usage is not linear, however.

Conclusions

Our novel tool MARV allows for RV analysis of multiple phenotypes in a computationally efficient and user-friendly manner. The data input formats and the command line interface familiar from widely-used GWAS software will offer researchers a quick setup for the analyses. Moreover, the feature of analysing all phenotype combinations within one run and the calculation of BIC to help in model selection will pave the way for rapid discoveries and novel insights into biology of complex traits.

Table 2 Computational time and peak memory usage of MARV by varying sample size, chromosomal size and number of phenotypes

Number of phenotypes (number of fitted models)	Chr 1 (249 Mbp)	Chr 22 (35 Mbp)
N = 2,405	h:min:s (memory)	h:min:s (memory)
2 (3)	4:06:54 (215 MB)	00:38:04 (260 MB)
4 (15)	3:51:23 (215 MB)	00:38:37 (260 MB)
8 (63)	5:07:11 (215 MB)	00:55:58 (260 MB)
N = 4,809	h:min:s (memory)	h:min:s (memory)
2 (3)	14:47:34 (780 MB)	02:26:00 (500 MB)
4 (15)	13:40:11 (780 MB)	02:26:10 (600 MB)
8 (63)	17:26:08 (780 MB)	03:03:00 (600 MB)

Methods

Statistical model

MARV is based on a so-called "reverse regression" approach, i.e. as compared to the standard GWAS in which the phenotype is the outcome and the genotype the predictor, this scenario is reversed in MARV. By using the genetic data as the outcome, we enable assessment of associations with multiple phenotypes simultaneously through the use of simple linear regression. While the "reverse regression" approach has been proposed for single genetic variants with the risk allele count or allele dosage being the outcome [36, 37], MARV uses a mutational load (burden) of risk alleles at RVS as the outcome. That is, the outcome is the proportion of RVs at which minor alleles are carried by individuals within a genomic region. This proportion is then modelled as a linear combination of *K* phenotypes. Mathematically, if r_i is the number of minor alleles at RVs and n_i is the total number of RVs, the model becomes:

$$r_i n_i^{-1} = \alpha + \boldsymbol{\beta}\mathbf{y}_\mathrm{i} + \varepsilon_\mathrm{i},$$

where $r_i n_i^{-1}$ is the proportion of minor alleles for *i*th individual, $\mathbf{y}_i$ is a vector of phenotype data for individual *i*, with corresponding regression coefficients $\boldsymbol{\beta} = [\beta_1,\ldots, \beta_K]$, and $\varepsilon_\mathrm{i} \sim \mathrm{MVN}(0,\sigma^2)$, σ^2 being a covariance matrix. Weighted linear regression is applied to allow for weighting by the number of successfully genotyped or imputed RVs within the region of interest. The significance testing is based on a likelihood ratio test which compares the weighted likelihoods of the fitted model against a null model where $\boldsymbol{\beta} = 0$. The likelihood ratio test statistic has an approximate χ2 distribution with *K* degrees of freedom.

The type I error rate and power of the method have been tested under various scenarios with simulated phenotype and genotype data, and the results from these analyses are described in detail elsewhere [19].

Availability and requirements

Project name: MARV

Project home page: https://github.com/ImperialStatGen/MARV

Operating system(s): UNIX

Programming language: C++

Other requirements: Standard Linux/UNIX build tools to compile the program.

License: BSD 3-Clause License

Any restrictions to use by non-academics: None

Additional files

Additional file 1: marv_chr1-5.result. Original results files outputted from MARV from the case study using NFBC1966 data. (RESULT 17825 kb)

Additional file 2: marv_chr6-10.result. Original results files outputted from MARV from the case study using NFBC1966 data. (RESULT 12614 kb)

Additional file 3: marv_chr11-15.result. Original results files outputted from MARV from the case study using NFBC1966 data. (RESULT 11495 kb)

Additional file 4: marv_chr16-22.result. Original results files outputted from MARV from the case study using NFBC1966 data. (RESULT 14097 kb)

Abbreviations

BIC: Bayesian information criterion; FI: Fasting insulin; GWAS: Genome-wide association studies; MAF: Minor allele frequency; MARV: Multi-phenotype analysis of rare variants; MPA: Multi-phenotype analysis; NFBC1966: Northern Finland Birth Cohort; RV: Rare variant; TG: Triglycerides; WHR: Waist-to-hip ratio

Acknowledgements

Northern Finland Birth Cohort (NFBC1966) would like to thank the late professor Paula Rantakallio (launch of NFBC1966), the participants in the 31 year study and the NFBC project center.

Funding

This work used the computing resources of the UK MEDical BIOinformatics partnership - aggregation, integration, visualisation and analysis of large, complex data (UK MED-BIO) which is supported by the Medical Research Council [grant number MR/L01632X/1]; the Imperial College High Performance Computing Service, URL: http://www.imperial.ac.uk/admin-services/ict/self-service/research-support/hpc/.

MK is funded by the European Commission under the Marie Curie Intra-European Fellowship (project MARVEL (WPGA-P48951)). APM is a Wellcome Trust Senior Fellow in Basic Biomedical Science (WT098017). IP was in part funded by the Elsie Widdowson Fellowship. NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 2/97, 8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 160/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231.

Availability of data and materials

The MARV software tool is freely available at URL: https://github.com/ImperialStatGen/MARV.

The Northern Finland Birth Cohort data which were used for the application of the developed tool are available upon collaboration and formal data request only, please see http://www.oulu.fi/nfbc/node/18136. The results from the case study are available as Additional files 1, 2, 3 and 4.

Authors' contributions

MK, RM, APM and JH wrote the MARV code and developed the software. MK, RM, APM and IP developed the methodology. MK performed the analyses and drafted the manuscript. JH tested the running time and memory usage. M.R.J. contributed to NFBC1966 data acquisition and study design. All authors read, edited and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The Ethics Committees of the University of Oulu and Northern Ostrobothnia Hospital District have approved the study. Individuals used for the analyses have provided written, informed consent.

Author details

[1]Department of Genomics of Common Disease, Imperial College London, London W12 0NN, UK. [2]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia. [3]Neuroepidemiology and Ageing (NEA) Research Unit, Imperial College London, London W6 8RP, UK. [4]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London W2 1PG, UK. [5]Center for Life Course Health Research, University of Oulu, 90014 Oulu, Finland. [6]Unit of Primary Care, Oulu University Hospital, 90220 Oulu, Finland. [7]Biocenter Oulu, University of Oulu, 90014 Oulu, Finland. [8]Department of Biostatistics, University of Liverpool, Liverpool L69 3BX, UK.

Received: 21 June 2016 Accepted: 6 February 2017
Published online: 16 February 2017

References

1. Schatz MC. Biological data sciences in genome research. Genome Res. 2015; 25:1417–22.
2. McVean GA, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Dinh H, Kovar C, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(V):56–65.
3. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JRB, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Brent Richards J, Greenwood CMT, Timpson NJ, Durbin R, Soranzo N, Bala S, Clapham P, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82–90.
4. The Haplotype Reference Consortium [http://www.haplotype-reference-consortium.org/home]. Accessed 8 Feb 2017.
5. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng H-F, UK10K Consortium, Gambaro G, Richards JB, Durbin R, Timpson NJ, Marchini J, Soranzo N. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun. 2015;6:8111.
6. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. Am J Hum Genet. 2014;95:5–23.
7. Galesloot TE, Van Steen K, Kiemeney LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. PLoS One. 2014;9:1–8.
8. Amos CI, Laing A. A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol. 1993;10:671–6.
9. Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. Am J Hum Genet. 1998;63:1190–201.
10. Banerjee S, Yandell BS, Yi NJ. Bayesian quantitative trait loci mapping for multiple traits. Genetics. 2008;179(August):2275–89.
11. Kim S, Xing EP. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 2009;5:e1000587.
12. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics. 1995;140:1111–27.
13. Shriner D. Moving toward system genetics through multiple trait analysis in genome-wide association studies. Front Genet. 2012;3(January):1.
14. Zhao J, Thalamuthu A. Gene-based multiple trait analysis for exome sequencing data. BMC Proc. 2011;5 Suppl 9:S75.

15. Marttinen P, Gillberg J, Havulinna A, Corander J, Kaski S. Genome-wide association studies with high-dimensional phenotypes. Stat Appl Genet Mol Biol. 2013;12:413–31.
16. Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, Xiong M, Wu CO, Fan R. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol. 2015;39:259–75.
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(September):559–75.
18. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39:906–13.
19. Kaakinen M, Mägi R, Fischer K, Heikkinen J, Järvelin M-R, Morris AP, Prokopenko I. A rare variant test for high-dimensional data. Eur J Hum Genet. 2017. Under revision.
20. Mägi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP. Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. Genet Epidemiol. 2012;36:785–96.
21. Rantakallio P. Groups at risk in low birth weight infants and perinatal mortality. Acta Paediatr Scand. 1969;193 Suppl 193:1+.
22. Scott R, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, Mägi R, Strawbridge RJ, Rehnberg E, Gustafsson S, Kanoni S, Rasmussen-Torvik LJ, Yengo L, Lecoeur C, Shungin D, Sanna S, Sidore C, Johnson PCD, Jukema JW, Johnson T, Mahajan A, Verweij N, Thorleifsson G, Hottenga J-J, Shah S, Smith AV, Sennblad B, Gieger C, Salo P, Perola M, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet. 2012;44:991–1005.
23. Sabatti C, Service SK, Hartikainen A-L, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Järvelin M-R, Freimer NB, Peltonen L. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet. 2009;41:35–46.
24. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC genome browser database: 2015 update. Nucleic Acids Res. 2015;43(November 2014):D670–81.
25. Ganna A, Salihovic S, Sundström J, Broeckling CD, Hedman ÅK, Magnusson PKE, Pedersen NL, Larsson A, Siegbahn A, Zilmer M, Prenni J, Ärnlöv J, Lind L, Fall T, Ingelsson E. Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. PLoS Genet. 2014;10: e1004801.
26. Feitosa MF, Wojczynski MK, Straka R, Kammerer CM, Lee JH, Kraja AT, Christensen K, Newman AB, Province MA, Borecki IB. Genetic analysis of long-lived families reveals novel variants influencing high density-lipoprotein cholesterol. Front Genet. 2014;5(June):159.
27. Major JM, Yu K, Wheeler W, Zhang H, Cornelis MC, Wright ME, Yeager M, Snyder K, Weinstein SJ, Mondul A, Eliassen H, Purdue M, Hazra A, McCarty CA, Hendrickson S, Virtamo J, Hunter D, Chanock S, Kraft P, Albanes D. Genome-wide association study identifies common variants associated with circulating vitamin E levels. Hum Mol Genet. 2011;20:3876–83.
28. Major JM, Yu K, Weinstein SJ, Berndt SI, Hyland PL, Yeager M, Chanock S, Albanes D. Genetic variants reflecting higher vitamin e status in men are associated with reduced risk of prostate cancer. J Nutr. 2014;144:729–33.
29. Cha S, Yu H, Park AY, Song KH. Effects of apolipoprotein A5 haplotypes on the ratio of triglyceride to high-density lipoprotein cholesterol and the risk for metabolic syndrome in Koreans. Lipids Health Dis. 2014;13:45.
30. Gaunt TR, Zabaneh D, Shah S, Guyatt A, Ladroue C, Kumari M, Drenos F, Shah T, Talmud PJ, Casas JP, Lowe G, Rumley A, Lawlor DA, Kivimaki M, Whittaker J, Hingorani AD, Humphries SE, Day IN. Gene-centric association signals for haemostasis and thrombosis traits identified with the HumanCVD BeadChip. Thromb Haemost. 2013;110:995–1003.
31. Grallert H, Dupuis J, Bis JC, Dehghan A, Barbalic M, Baumert J, Lu C, Smith NL, Uitterlinden AG, Roberts R, Khuseyinova N, Schnabel RB, Rice KM, Rivadeneira F, Hoogeveen RC, Fontes JD, Meisinger C, Keaney JF, Lemaitre R, Aulchenko YS, Vasan RS, Ellis S, Hazen SL, Van Duijn CM, Nelson JJ, März W, Schunkert H, McPherson RM, Stirnadel-Farrant H, Psaty BM, et al. Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: Meta-analysis of genome-wide association studies from five community-based studies. Eur Heart J. 2012;33(September 2010):238–51.
32. Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulcheko Y, Zhang W, Yuan X, Lim N. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. Atheroscler Thromb Vasc Biol. 2014;30:2264–76.
33. Suchindran S, Rivedal D, Guyton JR, Milledge T, Gao X, Benjamin A, Rowell J, Ginsburg GS, McCarthy JJ. Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study. PLoS Genet. 2010;6:e1000928.
34. Surakka I, Horikoshi M, Mägi R, Sarin A-P, Mahajan A, Lagou V, Marullo L, Ferreira T, Miraglio B, Timonen S, Kettunen J, Pirinen M, Karjalainen J, Thorleifsson G, Hägg S, Hottenga J-J, Isaacs A, Ladenvall C, Beekman M, Esko T, Ried JS, Nelson CP, Willenborg C, Gustafsson S, Westra H-J, Blades M, de Craen AJM, de Geus EJ, Deelen J, Grallert H, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015;47:589–97.
35. Tokoro F, Matsuoka R, Abe S, Arai M, Noda T, Watanabe S, Horibe H, Fujimaki T, Oguri M, Kato K, Minatoguchi S, Yamada Y. Association of a genetic variant of the ZPR1 zinc finger gene with type 2 diabetes mellitus. Biomed Rep. 2015;3:88–92.
36. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin M-R, Coin LJM. MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. PLoS One. 2012;7:e34861.
37. Mägi R, Suleimanov YV, Clarke GM, Kaakinen M, Fischer K, Prokopenko I, Morris AP. SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. BMC Bioinformatics. 2017. Accepted.