A rare-variant test for high-dimensional data

Marika Kaakinen¹, Reedik Mägi², Krista Fischer², Jani Heikkinen¹,³, Marjo-Riitta Järvelin⁴,⁵,⁶,⁷, Andrew P Morris⁸ and Inga Prokopenko*,¹

Genome-wide association studies have facilitated the discovery of thousands of loci for hundreds of phenotypes. However, the issue of missing heritability remains unsolved for most complex traits. Locus discovery could be enhanced with both improved power through multi-phenotype analysis (MPA) and use of a wider allele frequency range, including rare variants (RVs). MPA methods for single-variant association have been proposed, but given their low power for RVs, more efficient approaches are required. We propose multi-phenotype analysis of rare variants (MARV), a burden test-based method for RVs extended to the joint analysis of multiple phenotypes through a powerful reverse regression technique. Specifically, MARV models the proportion of RVs at which minor alleles are carried by individuals within a genomic region as a linear combination of multiple phenotypes, which can be both binary and continuous, and the method accommodates directly the genotyped and imputed data. The full model, including all phenotypes, is tested for association for discovery, and a more thorough dissection of the phenotype combinations for any set of RVs is also enabled. We show, via simulations, that the type I error rate is well controlled under various correlations between two continuous phenotypes, and that the method outperforms a univariate burden test in all considered scenarios. Application of MARV to 4876 individuals from the Northern Finland Birth Cohort 1966 for triglycerides, high- and low-density lipoprotein cholesterols highlights known loci with stronger signals of association than those observed in univariate RV analyses and suggests novel RV effects for these lipid traits.

European Journal of Human Genetics (2017) 25, 988–994; doi:10.1038/ejhg.2017.90; published online 24 May 2017

INTRODUCTION

In the past decade, thousands of loci for hundreds of phenotypes have been identified through genome-wide association studies (GWAS). Despite these discoveries, the issue of missing heritability¹ remains unsolved for most complex traits. While real-life analyses and power estimations suggest improvements for discovery with very large sample sizes, for many complex traits the numbers required are unrealistic. A cost-effective alternative for improved power is to jointly analyse multiple traits, by taking advantage of the correlation structure between them.²–⁶

A number of methods for multi-phenotype analysis (MPA) have been developed, including dimension reduction, and multivariate and graphical models.⁷ In genetics, MPA was first proposed for linkage studies,²,³,⁸,⁹ and recently, a plethora of methods for single-variant MPA of association has been suggested.¹⁰ MPA has been shown to provide a boost in power for locus discovery,²,⁵ as well as increased precision of parameter estimates.⁹ There is also a biological advantage in analysing correlated traits jointly, since detecting loci that affect a combination of phenotypes could provide suggestions of pleiotropic effects,⁷ that is, one locus affecting multiple phenotypes in parallel. Indeed, many of the identified loci from single-phenotype analyses overlap, especially for epidemiologically correlated traits, for example, glycaemic traits share up to a half of the loci with other cardiometabolic traits, including type 2 diabetes, lipids, measures of central obesity, height, blood pressure and hypertension.¹¹ This overlap suggests shared genetic architecture underlying at least a part of the observed epidemiological correlations. With joint analysis of multiple correlated phenotypes, the potentially shared genetic architecture would be better addressed by direct joint modelling. In addition, current technology allows for cost-effective deep phenotyping¹² such as the production of the metabolomics data including hundreds or even thousands of variables. These large-scale omics datasets will be necessary in understanding the exact relationship between genes and phenotypes. However, the great potential held by the omics data will only be achieved via integration and development of highly scalable computational and quantitative approaches.¹³ MPA methods will be particularly useful for the omics data in order to reduce dimensionality and avoid the penalties posed by correction for multiple testing.

Another way to improve power for locus discovery is to allow for a wider allele frequency range, including low-frequency and rare variants (minor allele frequency, MAF<5%, both denoted here by RVs). Development in technology to produce high-quality, low-cost sequencing data has had many positive implications for RV identification. Although whole-genome/exome sequencing is still not feasible at a large scale, that is, sequencing hundreds of thousands of individuals, accurate RV data can be produced with imputation based on dense reference panels such as the 1000 genomes project,¹⁴ the UK10K Project¹⁵ or the Haplotype Reference Consortium.¹⁶ For example, imputation to the UK10K reference panel including 7562 haplotypes, after re-phasing with SHAPEIT v2,¹⁷ yielded for variants with MAF as low as 0.1% an $r^2=0.5$ with genotyped variants.¹⁸ Combining these factors has led to an increased interest in RV

¹Department of Genomics of Common Disease, Imperial College London, London, UK; ²Estonian Genome Center, University of Tartu, Tartu, Estonia; ³Neuroepidemiology and Ageing (NEA) Research Unit, Imperial College London, London, UK; ⁴Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK; ⁵Center for Life Course Health Research, University of Oulu, Oulu, Finland; ⁶Unit of Primary Care, Oulu University Hospital, Oulu, Finland; ⁷Biocenter Oulu, University of Oulu, Oulu, Finland; ⁸Department of Biostatistics, University of Liverpool, Liverpool, UK; ⁹Correspondence: Dr I Prokopenko, Department of Genomics of Common Disease, Imperial College London, Du Cane Road, London, UK. Tel: +44 (0)207 594 6501; E-mail: i.prokopenko@imperial.ac.uk

Received 5 August 2016; revised 17 February 2017; accepted 28 March 2017; published online 24 May 2017
association analysis, resulting in the development of burden tests using collapsing techniques, variance-component tests and combinations of these two. We suggest integrating MPA with RV association analysis to further increase the power of locus discovery and to provide novel biological insights into complex trait genetics. Currently available methods for MPA are mostly tailored for single-variant associations, and are thus underpowered to detect RV associations. Some recently developed methods have addressed MPA for RVs, but they suffer, for example, from scalability limitations, their inability to combine continuous and binary phenotypes or to analyse different phenotype combinations, and finally, the lack of efficient computational tools (Figure 1). Here we introduce a burden test-based multi-phenotype analysis of RVs, MARV, which has several advantages over the few previously proposed methods that we are aware of: it is applicable to sequencing, imputed or genotyping data, allows for both binary and continuous phenotypes, provides association results for all phenotype combinations, including univariate tests within one run, and is implemented into a simple to use and computationally efficient software tool that can analyse millions of variants (Figure 1).

**MATERIALS AND METHODS**

**Multi-phenotype analysis of rare variants through reverse regression**

Our method is based on the RV burden test approach, in which RVs within a genomic region, defined by positional boundaries and/or annotation, are collapsed into one variable. Specifically, within a genomic region, we calculate the proportion of RVs at which an individual carries minor alleles: \( r_{in}^{-1} \), where \( r_i \) is the number of minor alleles at RVs and \( n_i \) is the total number of RVs. We then model this proportion as a linear combination of K phenotypes, that is, we use ‘reverse regression’ as compared to standard GWAS, with the genotype data as the outcome and the phenotypes as the predictors, in the same way as in the MultiPhen approach for common variants. Thus, the model becomes

\[
E(r_{in}^{-1}) = \alpha + \beta y_i,
\]

where \( r_{in}^{-1} \) is the proportion of minor alleles for ith individual, \( y_i \) represents the phenotype data for individual i, with corresponding regression coefficients \( \beta = (\beta_1, \ldots, \beta_K) \). The modelling is performed via weighted linear regression with the following weights: the proportion of successfully genotyped or imputed RVs within the region of interest. Parameter estimates are obtained via least squares estimation. Then, a likelihood ratio test is constructed by comparing the weighted log likelihoods of the fitted model against a null model where \( \beta = 0 \). The test statistic has an approximate \( \chi^2 \) distribution with K degrees of freedom.
The method described has been implemented in a freely available software tool MARV, which is available at https://github.com/ImperialStatGen/MARV.

The methodology is flexible such that it can accommodate both quantitative and binary phenotypes, as well as directly genotyped and imputed RVs. For imputed RVs, instead of using the 0/1 indicator for the absence/presence of minor allele, we use the posterior probability that an individual is heterozygous or rare homozygous. For discovery purposes, the full model, including all the phenotypes is fitted. However, to allow further investigation of the loci reaching genome-wide significance after correction for multiple testing to take into account the number of regions tested within the analysis, we have implemented in the MARV software the possibility to analyse all phenotype combinations. For model selection purposes, MARV further calculates the Bayesian information criterion, $BIC = -2 \ln(L) + K \ln(N)$, where $L$ is the estimated likelihood of the model, $K$ is the number of parameters and $N$ is the number of observations.

Genotype simulations

To evaluate the type I error rate and power of our method, we simulated genetic variants by using a model that results in an allele spectrum similar to that observed in the European population. Specifically, we used the forward simulation software tool ForSim and the hybrid model for population history proposed previously. In brief, the ancestral population size was assumed to be 8100 for 50000 generations, followed by a bottleneck population size of 2000. After this time the population expands with an exponential growth rate of 1.29% for over 370 generations, resulting in a modern effective population size of 227 650. With a maximum number of offspring set to two, the simulation resulted in a population of about 500 000 individuals. The mutation rate was assumed to be $2.0 \times 10^{-8}$ bp per generation. This hybrid model has features of two published demographic histories and has been shown to recapitulate the number and frequency distributions of both rare and common synonymous sites, as well as empirically observed patterns of linkage disequilibrium between common variants.

We modelled loci as a series of 8 exons and 7 introns (alternating), with exons of length 300 bp and introns of length 3 kb such that the total coding length is 2.4 kb and the total transcript length is 23.4 kb, which corresponds to two published demographic histories and has been shown to recapitulate the number and frequency distributions of both rare and common synonymous sites, as well as empirically observed patterns of linkage disequilibrium between common variants.

Phenotype simulations

We simulated two continuous phenotypes from a multivariate normal distribution, such that

$$Y = [y_1, y_2]^T = \beta_0 + \cdots + \beta_k y_k + \epsilon$$

where $y_k$ is the causal variant and $\beta_k$ the effect for each causal variant, $k = 1, \ldots, 146$, and $\epsilon \sim \text{MVN}(0, \Sigma_f)$, $\Sigma_f = \text{cov}(y_1, y_2) = \begin{bmatrix} \sigma_1^2 & \rho_{12} \\ \rho_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, and $\rho \in [-0.9, 0.9]$. We used the function mvrnorm in the statistical software R. The genetic effects $\beta_k$ were considered under three scenarios, including (1) no effect (for type I error rate assessment), that is, $\beta_k = 0$; (2) all causal genetic variants having a trait-increasing effect ($\beta_k = 0.1$), and (3) half of the causal variants having a trait-increasing effect ($\beta_k = 0.1$) and half having a trait-decreasing effect ($\beta_k = -0.1$). Scenarios (2) and (3) were further considered in situations when there are effects on (a) both phenotypes in the same direction with same magnitudes, (b) both phenotypes in opposing directions with same magnitudes, (c) both phenotypes in the same direction with different magnitudes (the effect on phenotype two is half of that for phenotype one), and (d) only on one of the phenotypes.

Type I error rate and power estimation

We estimated the type I error rate and power of our method using the full model by calculating the proportion of the 10 000 replicate analyses, in which $P < 0.05$ for the test of association. Since we analysed only the full model and one genomic region, no correction for multiple testing was applied.

We compared the power of MARV to that of a recently published multi-phenotype RV method GAMuT, which is based on a non-parametric distance-covariance test. This method requires a phenotype similarity matrix, either by using a projection matrix or by using user-selected linear kernel functions. For our estimates, we constructed the similarity projection matrix. The genotypic similarity matrix was calculated assuming a linear kernel for genotypes, where weights are functions of MAF, following the notation from Broadaway et al. The method then constructs the test and provides the $P$-value as the output. Similarly to MARV, we estimated the power of GAMuT by calculating the proportion of the 10 000 replicate analyses in which $P < 0.05$ for the test of association.

In addition, we compared the power of MARV over a univariate RV test by calculating associations between the simulated RVs and the two phenotypes separately using the GRANVIL software. To assess power, we calculated the proportion of tests for which $P < 0.025$, where the level of significance is Bonferroni corrected for two tests performed.

Experimental setup

To examine our method with a real dataset, we analysed the data for triglycerides (TG), low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol from individuals belonging to the Northern Finland Birth Cohort 1986 (NBBC1986) and having participated in the 31-year clinical examination. The cohort covers over 96% of all births in the two northernmost provinces of Finland in 1966 (N = 12 068 live-born children). The Ethics Committees of the University of Oulu and Northern Ostrobothnia Hospital District have approved the study, and the individuals used for the analyses have provided written informed consent.

At the clinical examination, blood samples were drawn after overnight fasting. Samples were stored at −70°C until analysed. Enzymatic assays of fasting HDL cholesterol and TG were measured using Hitachi 911 Clinical Chemistry Analyzer and commercial reagents (Boehringer, Mannheim, Germany) in the accredited laboratory of Oulu University Hospital. LDL was obtained using a previously described method. DNA was extracted and genome-wide genotyping was performed with the Illumina HumanCNV370-DUO Analysis BeadChip platform at the Broad Institute, USA. The Beadstudio algorithm was used for calling the genotypes. Detailed genotyping and sample quality control (QC) of the first set of data have been reported previously. More samples were genotyped afterwards, and after the genotyping data QC, there were 5402 subjects and 324 896 SNPs available for analysis. SNPs were imputed to the 1000 Genomes all ancestries reference panel (March 2012) with IMPUTE2, resulting in ~38M SNPs for analysis. Of these, 7 396 876 markers were non-monomorphic with MAF < 5% and good imputation quality (info > 0.4).

For the phenotypes, we excluded non-fasting individuals and those known to be on lipid-lowering medication. Residuals of TG, LDL and HDL were calculated by adjusting for sex, body mass index and the first three principal components (PCs) derived from the genetic data to control for potential population structure. The residuals were further inverse-normal transformed; however, it is worth noticing that a transformation to reduce skewness is not a prerequisite for the software as the phenotypes are not the outcomes of interest here. GWAS and the phenotype data were available for 4876 individuals.

The multi-phenotype genome-wide RV analysis using the transformed residuals was performed with MARV using the method "expected", that is, we used the genotype dosages coming from imputation. To define gene regions across the genome, we used the gene list from the University of California Santa Cruz (UCSC, NCBI genome sequence build 37, hg19). We applied a Bonferroni correction for 30 000 genes resulting in a level of significance of $1.67 \times 10^{-6}$. We analysed all variants irrespective of their annotation across autosomal chromosomes using the following thresholds: MAF < 5% and imputation quality > 0.4.
RESULTS

Type I error rate and power

The simulation studies on 5000 individuals showed a good control of type I error rate under various correlation structures between the two continuous phenotypes (Figure 2). Supplementary Figure S1 demonstrates that the control is maintained with a substantially smaller sample size ($N=1000$). Numerical results for both sample sizes are provided in Supplementary Table S1.

In all the tested scenarios, when all genetic effects are trait-increasing, power is almost always higher than that of univariate analyses or the kernel-based multi-phenotype test GAMuT (Figures 3a–d and Supplementary Table S2). When the genetic effects are in the same direction and of the same magnitude for the two phenotypes (Figure 3a), power is highest when the correlation between the phenotypes is highly negative. The opposite is observed, that is, power is highest with high positive correlation between the phenotypes when the genetic effects are in the opposite directions but of the same magnitude (Figure 3b). When there are genetic effects in the same direction, but with different magnitude (half of the effect size of phenotype 1 in phenotype 2), the pattern of power follows that of the scenario in Figure 3a, that is, power is higher with negative correlation between the phenotypes (Figure 3c). However, the power decreases at a faster rate compared to 3a) and increases again when the correlation is equal to or higher than 0.5. Finally, when the genetic variants affect only one of the correlated phenotypes, power is increased both when the phenotypes are highly positively or highly negatively correlated (Figure 3d). Even with no correlation between the phenotypes, there is an improvement in power over univariate analyses, which is due to the avoidance of correction for multiple testing with our multivariate model, despite the requirement of an additional degree of freedom. Results from simulations using a sample size of 1000 individuals are shown in Supplementary Figure S2. Even with a smaller sample size, joint analysis of correlated phenotypes offers an improvement in power over the univariate analysis of 5000 individuals when the correlation between the phenotypes is high enough (over 0.5 or less than −0.5).

When we considered the above mentioned four scenarios (Figures 3a–d) under the assumption that half of the genetic variants are trait-increasing and half trait-decreasing, we see similar patterns of power as for the situation of all variants being trait-increasing, but in general, the power is lower (Figures 3e–h and Supplementary Table S2). In all cases, MARV performed better than univariate analyses. The power of GAMuT was only very modestly increased over that of MARV with both, trait-increasing and -decreasing directions of genetic effects on highly correlated phenotypes.

![Figure 2](image)

Figure 2 Estimated type I error rate with 95 % confidence interval (CI) of the MARV method with $N=5000$ and varying correlation between two continuous phenotypes. The following correlations were evaluated: −0.9, −0.5, −0.3, −0.1, 0, 0.1, 0.3, 0.5, 0.9.

![Figure 3](image)

Figure 3 Statistical power of the MARV method with $N=5000$ and varying correlation between two continuous phenotypes. (a–d) All genetic effects are trait-increasing. (e–h) Half of the genetic effects are trait-increasing, half trait-decreasing. (a,e) Effects on both phenotypes, same direction, same magnitude. (b,f) Effects on both phenotypes, opposite direction, same magnitude. (c,g) Effects on both phenotypes, same direction, different magnitude (effect on phenotype 2 is half of that on phenotype 1). (d,h) Effects on one phenotype only. Solid, black line: MARV; dotted, magenta line: GAMuT; dashed, grey line: univariate analysis (GRANVIL). The following correlations were evaluated: −0.9, −0.5, −0.3, −0.1, 0, 0.1, 0.3, 0.5, 0.9.
The three selected phenotypes, TG, HDL and LDL, were modestly correlated with each other ($R_{TG,HDL} = -0.32$, $R_{TG,LDL} = 0.31$, $R_{HDL,LDL} = -0.19$). The MPA of RVs for the three phenotypes revealed genome-wide significant associations ($P < 1.67 \times 10^{-6}$, Methods) covering five gene regions: SARS, CELSR2, APOA5, FAM63B and APOE (Figures 4a and b; Supplementary Tables S3 and S4). 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 analysis. Additionally, the BIC provided by MARV for each sub model served for selection of the phenotype combination providing the best fit. At SARS, APOA5 and APOE, the univariate models provided the best fit (Figure 4c; Supplementary Table S3). At CELSR2 and FAM63B, the best fit was coming from the combination of LDL+HDL and TG +HDL, respectively, and only the full and best models reached genome-wide significance, indicating that the associations would have been missed in univariate analyses. Both of these loci are also highly enriched for RVs, with 87 variants with MAF $< 5\%$ being included for CELSR2 analysis and 284 for FAM63B, while for the other associated genes the rare-variant count ranged between 3 and 19 (Supplementary Tables S3 and S4).

**DISCUSSION**

We propose a novel flexible method, MARV, for MPA of RVs across the genome applicable to the sequencing, imputed and genotyping data in unrelated individuals. This method using the powerful reverse regression approach should enable novel discoveries based on the large-scale genomic data with high DNA-variant density. We show with simulations that our method has a good control of type I error rate under various scenarios, even with a sample size as small as 1000 individuals. The power of our method always exceeds that of a univariate RV burden test under the tested scenarios, and the patterns
mirror those presented for a common-variant multi-phenotype method. Compared to a kernel-based multi-phenotype RV test, the power of MARV is notably increased when all the genetic effects are in the same direction, and is similar when half of the effects are protective and half are deleterious. The gain in power is observed due to the inclusion of correlated phenotypes in one model and thus, correction for multiple testing is avoided, given joint phenotype modelling. This feature is important for the efficient analysis of the large-scale high-dimensional omics data. For the univariate analyses, we applied the commonly used Bonferroni correction to account for the power of the univariate method. We note that the Bonferroni correction applied may have been too stringent, especially when the phenotypes are highly correlated, but this approach is standard in GWAS and is widely applied to such traits. With the univariate method, we applied the commonly used Bonferroni correction to account for the power of the univariate burden test. Similarly, we expect that power could be reduced in the presence of only a small number of causal variants. Combined burden and variance-component tests, such as SKAT-O, are claimed to be more robust than burden tests in the above mentioned scenarios. However, when we compared the power of MARV to a kernel-based multi-phenotype RV test GAMuT in the presence of both trait-increasing and trait-decreasing effects, the power of MARV is similar to that of GAMuT in all tested scenarios.

In summary, we have developed a powerful method to facilitate the discovery of RV genetic effects within the multi-phenotype framework. The method has been implemented into a freely available, easy to use software tool MARV, and should allow wide and rapid implementation in the analytical pipelines for large-scale high-dimensional datasets, therefore paving the way for novel genomic discoveries.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

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/. This project was supported by the European Commission under the Marie Curie Intra-European Fellowship (project MARVEL (PIEF-GA-2013-626641)). IP was in part funded by the Elsie Widdowson Fellowship, the Wellcome Trust Seed Award in Science (WT205915) and the European Union’s Horizon 2020 research and innovation programme (DYNAnet, project number 633595). APM is a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant number WT098017). Northern Finland Birth Cohort (NFBC1966) thank the late Professor Paula Rantakallio (launch of NFBC1966), the participants in the 31 years study and the NFBC project center. NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 297/8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231.

European Journal of Human Genetics

M Kaakinen et al

993
