European Journal of Human Genetics (2017) 25, 988–994
Official journal of The European Society of Human Genetics
www.nature.com/ejhg

EJHG Open

ARTICLE

A rare-variant test for high-dimensional data

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]

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 heritability[1] 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 phenotyping[12] 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 Project[15] 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

[1]Department of Genomics of Common Disease, Imperial College London, London, UK; [2]Estonian Genome Center, University of Tartu, Tartu, Estonia; [3]Neuroepidemiology and Ageing (NEA) Research Unit, Imperial College London, London, UK; [4]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK; [5]Center for Life Course Health Research, University of Oulu, Oulu, Finland; [6]Unit of Primary Care, Oulu University Hospital, Oulu, Finland; [7]Biocenter Oulu, University of Oulu, Oulu, Finland; [8]Department of Biostatistics, University of Liverpool, Liverpool, UK
*Correspondence: 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

PUBLISHED

RARE VARIANT SINGLE PHENOTYPE METHODS
(Review by Lee *et al.* 2014)

- **Burden tests**, e.g. CMC method, WSS
- **Adaptive burden tests**, e.g. KBAC method
- **Variance-component tests**, e.g. SKAT
- **Combined tests**, e.g. SKAT-O, Fisher method
- **EC test**

COMMON VARIANT MULTI-PHENOTYPE METHODS
(Comparison by Galesloot *et al.* 2014)

- **Direct methods**, e.g. MV-SNPTEST, MultiPhen, MV-BIMBAM
- **Indirect methods**, e.g. MV-PLINK, PCHAT, UV-PCA
- **Univariate-based methods**, e.g. TATES, UM-MA

RARE VARIANT MULTI-PHENOTYPE METHODS

Previously proposed methods

Method	Advantages	Disadvantages	Reference
Weighted-sum collapsing method with reduced univariate phenotype	❖ binary and continuous phenotypes ❖ next-generation sequencing data	❖ does not distinguish between common and rare variants ❖ method tested on ~24k SNPs only ❖ interpretation of affected phenotypes not straightforward when using a reduced phenotype ❖ no software tool available	Zhao and Thalamuthu 2011
Bayesian reduced rank regression –based method	❖ tens of phenotypes ❖ freely available R code	❖ computationally exhaustive ❖ not possible to directly identify affected phenotypes	Marttinen *et al.* 2014
The approximate F-distribution test	❖ freely available R code	❖ quantitative traits only ❖ method tested on ~54k variants only ❖ separate analyses needed for different phenotype combinations	Wang *et al.* 2015
A non-parametric distance-covariance based test, Gene Association with Multiple Traits, GAMuT	❖ tens of phenotypes, continuous and binary ❖ adjustment for covariates ❖ genome-wide data ❖ freely available R code	❖ not possible to directly identify affected phenotypes	Broadaway *et al.* 2016
A linear mixed model –based Multivariate Rare-Variant Association Test, MURAT	❖ tens of phenotypes ❖ genome-wide data ❖ computationally efficient ❖ freely available R code	❖ quantitative traits only ❖ not possible to directly identify affected phenotypes	Sun *et al.* 2016

NOVEL

Multi-phenotype Analysis of Rare Variants, MARV

Method	Advantages	Disadvantages	Reference
A burden test –based multi-phenotype method, MARV	❖ binary and continuous phenotypes ❖ genome-wide data, directly genotyped and imputed ❖ all phenotype combinations analysed, model selection based on BIC ❖ direct interpretation of affected phenotypes ❖ freely available, computationally efficient command line tool	❖ loss of power related to burden test, i.e. in the presence of both phenotype-increasing and phenotype-decreasing variants or a small fraction of causal variants	Kaakinen *et al.* 2017

Figure 1 Comparison of MARV with previously proposed RV multi-phenotype association analysis methods. Upper blocks: Established rare-variant single-phenotype methods[19] and common-variant multi-phenotype methods[10] based on the individual level data. Lower block: Previously proposed RV multiple-phenotype methods[20–24] versus our proposed MARV method.

association analysis, resulting in the development of burden tests using collapsing techniques, variance-component tests and combinations of these two.[19]

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,[20–24] 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,[25,26] in which RVs within a genomic region, defined by positional boundaries and/or annotation, for example, are collapsed into one variable. Specifically, within a genomic region, we calculate the proportion of RVs at which an individual i carries minor alleles: $r_i n_i^{-1}$, 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.[27] Thus, the model becomes

$$\mathrm{E}(r_i n_i^{-1}) = \alpha + \boldsymbol{\beta}\mathbf{y}_i,$$

where $r_i n_i^{-1}$ is the proportion of minor alleles for ith individual, $\mathbf{y}_i$ represents the phenotype data for individual i, with corresponding regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$. 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 $\boldsymbol{\beta} = 0$. The test statistic has an approximate χ^2 distribution with K degrees of freedom.

The method described has been implemented in a freely available software tool MARV,[28] 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, $BIC = -2 \bullet \ln(L) + K \bullet \ln(N)$, 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.[29] Specifically, we used the forward simulation software tool ForSim[30] and the hybrid model for population history proposed previously.[29] In brief, the ancestral population size was assumed to be 8100 for 50 000 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 2.0×10^{-8} bp per generation. This hybrid model has features of two published demographic histories[31,32] 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.[29]

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 an "average" protein-coding gene from the RefSeq database.[29] Around each transcript, we also simulated 100 kb of neutral genomic target flanking both sides of the gene.[29] The simulation resulted in 16 616 SNPs, of which 96% have a MAF $<5\%$. We randomly sampled SNPs as causal irrespective of their functional impact such that the MAF of any individual causal SNP was $<2\%$ and the sum of the MAFs did not exceed 5%.[33] This resulted in 146 causal SNPs that were used for all the tested scenarios. For the selected SNPs, we randomly sampled individuals from this final population to achieve 10 000 replicates of studies of 1000 and 5000 individuals.

Phenotype simulations

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

$$Y = [y_1 y_2]^T = \beta_1 g_1 + \ldots + \beta_n g_n + \varepsilon$$

where g_k is the causal variant and β_k the effect for each causal variant, $k = 1, \ldots, 146$, and $\varepsilon \sim \mathrm{MVN}(0, \Sigma_Y)$, $\Sigma_Y = \mathrm{cov}(y_1, y_2) = \begin{bmatrix} \sigma_1^2 & \rho_{12} \\ \rho_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix}$, and $\rho_{12} \in [-0.9, 0.9]$. We used the function mvrnorm in the statistical software R.[34] The genetic effects β_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,[23] 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*.[23] 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.[26] 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 1966 (NFBC1966) 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).[35] 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.[36] 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.[37] 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,[38] 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).[39] We applied a Bonferroni correction for 30 000 genes resulting in a level of significance of 1.67×10^{-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.

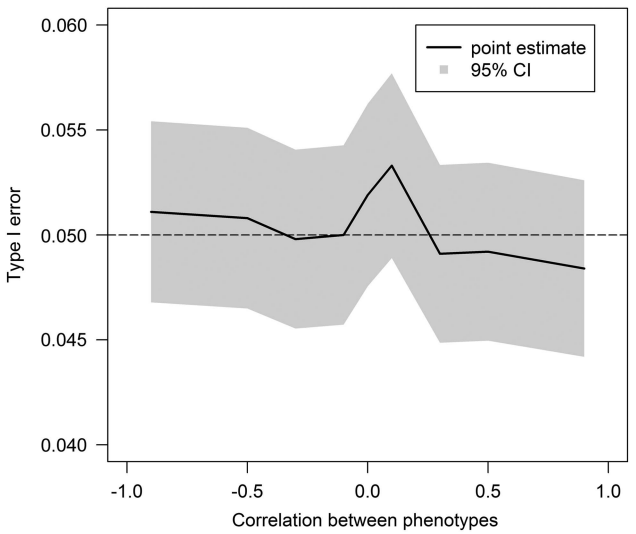

Figure 2 Estimated type I error rate with 95 % confidence interval (CI) of the MARV method with $N=5000$ and varying correlation between two continuous phenotypes. The following correlations were evaluated: −0.9, −0.5, −0.3, −0.1, 0, 0.1, 0.3, 0.5, 0.9.

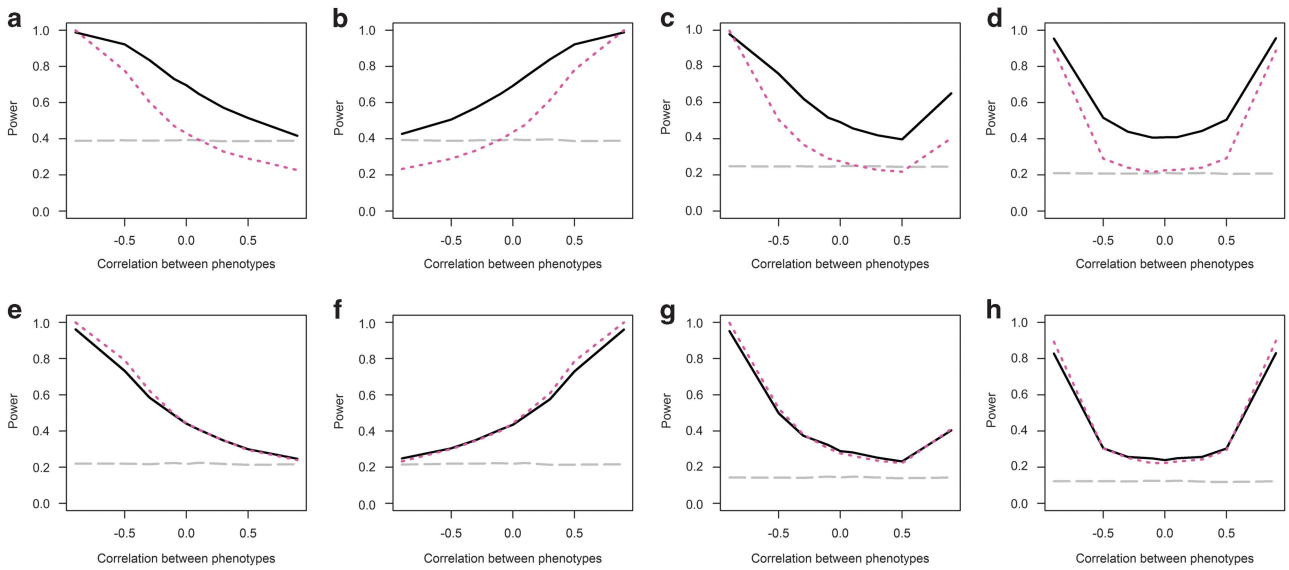

Figure 3 Statistical power of the MARV method with $N=5000$ and varying correlation between two continuous phenotypes. (**a**–**d**) All genetic effects are trait-increasing. (**e**–**h**) Half of the genetic effects are trait-increasing, half trait-decreasing. (**a**,**e**) Effects on both phenotypes, 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.

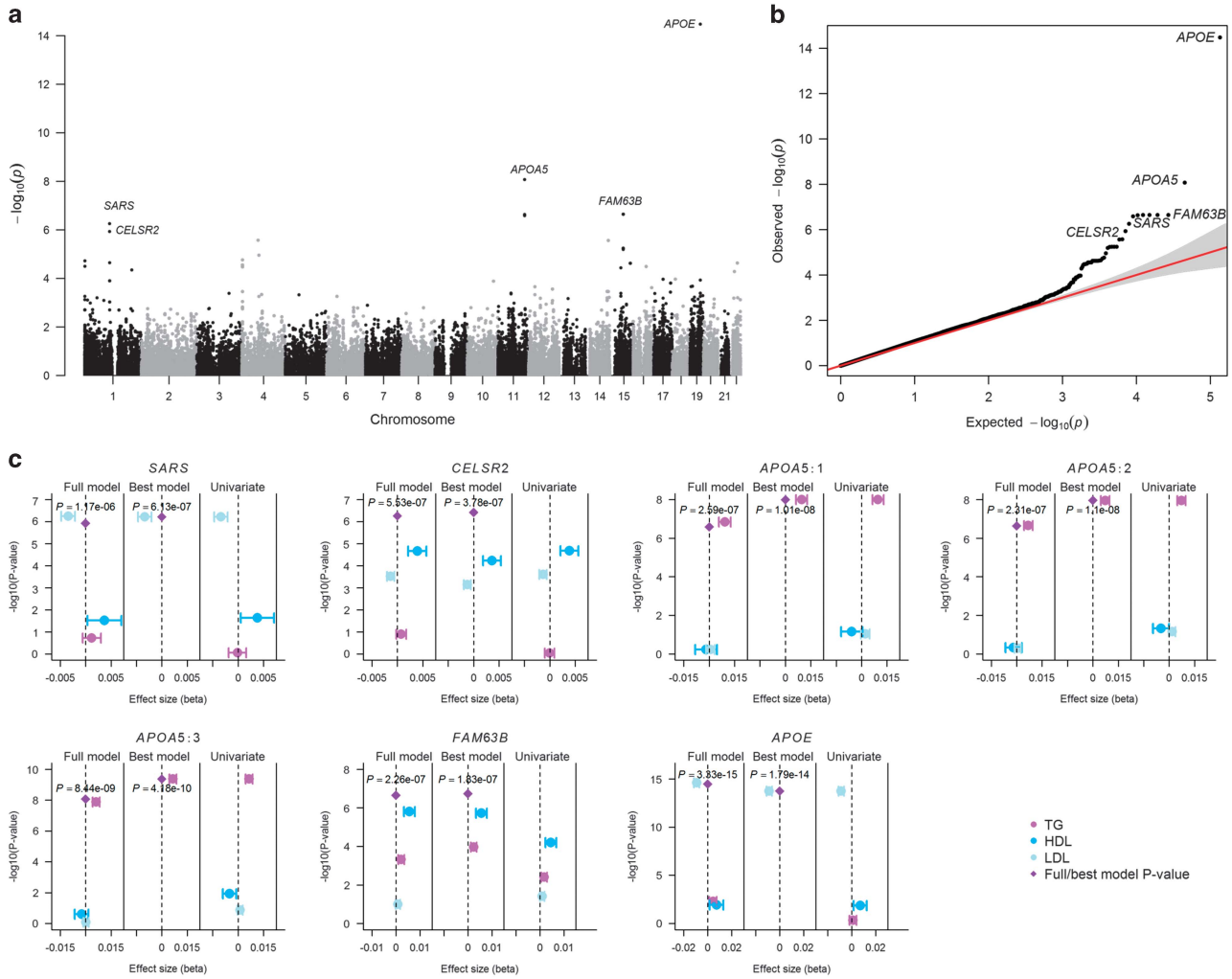

Figure 4 Genome-wide association analysis results from MARV for triglycerides, high-density lipoprotein and low-density lipoprotein cholesterols in the NFBC1966. (**a**) Manhattan plot for the full model statistical significance. Genes reaching statistical significance ($P<1.67\times10^{-6}$) are annotated. (**b**) QQ-plot of the full model association *P*-values against the expected *P*-values. Note that at some of the loci, different gene transcripts resulted in exactly the same association result. Such results show as a horizontal line of dots in the figure. (**c**) Effect sizes with their 95% confidence intervals of triglycerides, high-density lipoprotein and low-density lipoprotein cholesterols plotted against their statistical significance for the loci reaching genome-wide significance. In each figure, the panel on the left shows the results from the full model, the middle panel shows them from the best model based on Bayesian Information Criterion and the right panel illustrates results from univariate models. For *APOA5*, statistically significant associations were detected for three different transcripts.

Real data example

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 ($P<1.67\times10^{-6}$, 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.[27] 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 two tests (one for each phenotype) performed. This notably decreased 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[40] and is widely applied to such traits.[41]

With application of MARV to TG, HDL and LDL phenotypes, we were able to detect genome-wide significant associations with RVs even with a sample size as modest as 4876 individuals. We found evidence for novel RV associations at *CELSR2*, *SARS* and *FAM63B,* of which common variation at *CELSR2* has been previously associated with blood lipids,[42] and at *SARS/CELSR2* with type 2 diabetes,[43] while *FAM63B* has no reported links with blood lipid levels. We identified two additional well-established lipid genes, *APOA5* and *APOE*.[42] A recent study reported RV associations at *APOE* with lipid levels,[44] and another study associated RVs at *APOA5* with myocardial infarction.[45] Our findings are in line with previous reports of variation at *APOA5* and *APOE* mainly affecting TG and LDL levels, respectively;[42] however, we observed RV association at *CELSR2* with both HDL and LDL cholesterol levels, thus extending previous reports on common variation at this locus about primary effects on LDL.[42] The novel *FAM63B* has previously been associated with schizophrenia through methylation patterns,[46] making this locus an interesting biological candidate for lipid metabolism and worth further investigation, as metabolic disturbances often co-occur with psychotic illnesses.[47] Importantly, our analyses showed that the associations at this locus and at *CELSR2* would not have been observed in univariate analyses, whereas RV MPA had sufficient power to detect these signals. We also observed that MARV was able to detect effects from genes with widely varying sizes and number of RVs (*APOA5* with 2.5 Mbp and 3 RVs vs *FAM63B* with 86 Mbp and 284 RVs). It remains of future work, however, to formally assess the power of MARV under properties such as varying sets of causal variants, different gene size, mutation rate, haplotype length or degree of linkage disequilibrium between causal variants.

The simulated scenarios presented here are for two continuous phenotypes, but the method is easily applicable to a large number of phenotypes, as the real data example showed, and to a mixture of binary and continuous phenotypes. This is due to the methodological framework used, for example, reverse regression, in which the phenotypes are treated as predictors instead of outcomes, allowing for an easy incorporation of variables with varying properties into the model, as in any other linear regression analysis. Further, we have presented an example using a gene region-based association analyses, but the method and software are flexible and allow for analyses of any regions defined by the user based on start and end positions, for example, region sets, including those encoding active enhancers from analyses of human regulome.[48] MARV could also be adapted to focus on only the most deleterious variants, rather than all of those mapping to a gene region via extracting or excluding specific markers.

There are limitations to our method and the current version of the software that warrant discussion. MARV is based on a burden test, which has been shown to be a powerful method for RV association testing when a large proportion of variants are causal and their effects are in the same direction.[19] However, in the presence of both trait-increasing and trait-decreasing effects, or of only a small proportion of causal variants amongst those tested, there is a loss in power.[19] As MARV is a burden test, it also suffers from some loss of power when both trait-increasing and trait-decreasing effects impact phenotypes, as shown in Figure 3. However, the power is still always higher compared to the univariate burden test. Similarly, we expect that power could be reduced in the presence of only a small number of causal variants.[49] Combined burden and variance-component tests, such as SKAT-O, are claimed to be more robust than burden tests in the above mentioned scenarios.[19] 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-2013-626461)). 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 (DYNAhealth, 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. 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.

1 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
2 Amos CI, Laing A: A comparison of univariate and multivariate tests for genetic linkage. *Genet Epidemiol* 1993; **10**: 671–676.
3 Allison DB, Thiel BSt, 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**: 1190–1201.
4 Banerjee S, Yandell BS, Yi NJ: Bayesian quantitative trait loci mapping for multiple traits. *Genetics* 2008; **179**: 2275–2289.
5 Kim S, Xing EP: Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet* 2009; **5**: e1000587.
6 Ferreira MAR, Purcell SM: A multivariate test of association. *Bioinformatics* 2009; **25**: 132–133.
7 Shriner D: Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front Genet* 2012; **3**: 1.
8 Almasy L, Dyer TD, Blangero J: Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genet Epidemiol* 1997; **14**: 953–958.

9 Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 1995; **140**: 1111–1127.
10 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.
11 Marullo L, El-Sayed Moustafa JS, Prokopenko I: Insights into the genetic susceptibility to type 2 diabetes from genome-wide association studies of glycaemic traits. *Curr Diab Rep* 2014; **14**: 551.
12 Delude D: Deep phenotyping: The details of disease. *Nature* 2015; **527**: S14–S15.
13 Schatz MC: Biological data sciences in genome research. *Genome Res* 2015; **25**: 1417–1422.
14 McVean GA, Altshuler DM, Durbin RM *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
15 Walter K, Min JL, Huang J *et al*: The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
16 Marchini J, Abecasis GR, Durbin RM: The Haplotype Reference Consortium 2014. Available at: http://www.haplotype-reference-consortium.org/home.
17 Delaneau O, Marchini J, Zagury J-F: A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011; **9**: 179–181.
18 Huang J, Howie B, McCarthy S *et al*: Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015; **6**: 8111.
19 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.
20 Zhao J, Thalamuthu A: Gene-based multiple trait analysis for exome sequencing data. *BMC Proc* 2011; **5**: S75.
21 Wang Y, Liu A, Mills JL *et al*: Pleiotropy Analysis of Quantitative Traits at Gene Level by Multivariate Functional Linear Models. *Genet Epidemiol* 2015; **39**: 259–275.
22 Marttinen P, Pirinen M, Sarin AP *et al*: Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics* 2014; **30**: 2026–2034.
23 Broadaway KA, Cutler DJ, Duncan R *et al*: A Statistical approach for testing cross-phenotype effects of rare variants. *Am J Hum Genet* 2016; **98**: 525–540.
24 Sun J, Oualkacha K, Forgetta V *et al*: A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *Eur J Hum Genet* 2016; 1–8.
25 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010; **34**: 188–193.
26 Mägi R, Kumar A, Morris AP: Assessing the impact of missing genotype data in rare variant association analysis. *BMC Proc* 2011; **5**: S107.
27 O'Reilly PF, Hoggart CJ, Pomyen Y *et al*: MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* 2012; **7**: e34861.
28 Kaakinen M, Mägi R, Fischer K *et al*: MARV: a tool for genome-wide multi-phenotype analysis of rare variants. *BMC Bioinformatics* 2017; **18**: 110.
29 Agarwala V, Flannick J, Sunyaev S, Altshuler D: Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet* 2013; **45**: 1418–1427.
30 Lambert BW, Terwilliger JD, Weiss KM: ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics* 2008; **24**: 1821–1822.
31 Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci USA* 2009; **106**: 3871–3876.
32 Gravel S, Henn BM, Gutenkunst RN *et al*: Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci* 2011; **108**: 11983–11988.
33 Tachmazidou I, Morris A, Zeggini E: Rare variant association testing for next-generation sequencing data via hierarchical clustering. *Hum Hered* 2013; **74**: 165–171.
34 R Core Team: R: A language and environment for statistical computing; 2014. Available at: http://www.r-project.org/.
35 Rantakallio P: Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand* 1969; **193**: 1+.
36 Wieland H, Seidel D: A simple specific method for precipitation of low density lipoproteins. *J Lipid Res* 1983; **24**: 904–909.
37 Sabatti C, Service SK, Hartikainen A-L *et al*: Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 2009; **41**: 35–46.
38 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
39 Rosenbloom KR, Armstrong J, Barber GP *et al*: The UCSC genome browser database: 2015 update. *Nucleic Acids Res* 2015; **43**: D670–D681.
40 Balding DJ: A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006; **7**: 781–791.
41 Suhre K, Shin S-Y, Petersen A-K *et al*: Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2011; **477**: 54–60.
42 Willer CJ, Schmidt EM, Sengupta S *et al*: Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; **45**: 1274–1283.
43 Fontaine-Bisson B, Renstrom F, Rolandsson O *et al*: Evaluating the discriminative power of multi-trait genetic risk scores for type 2 diabetes in a northern Swedish population. *Diabetologia* 2010; **53**: 2155–2162.
44 Surakka I, Horikoshi M, Mägi R *et al*: The impact of low-frequency and rare variants on lipid levels. *Nat Genet* 2015; **47**: 589–597.
45 Do R, Stitziel NO, Won H-H *et al*: Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2015; **518**: 102–106.
46 Aberg KA, McClay JL, Nerella S *et al*: Methylome-wide association study of schizophrenia. *JAMA Psychiatry* 2014; **71**: 255.
47 Oresic M: Obesity and psychotic disorders: uncovering common mechanisms through metabolomics. *Dis Model Mech* 2012; **5**: 614–620.
48 Pasquali L, Gaulton KJ, Rodríguez-Seguí SA *et al*: Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 2014; **46**: 136–143.
49 Moutsianas L, Agarwala V, Fuchsberger C *et al*: The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet* 2015; **11**: e1005165.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)