Bioinformatics, 31(16), 2015, 2754–2756
doi: 10.1093/bioinformatics/btv222
Advance Access Publication Date: 22 April 2015
Applications Note

OXFORD

Genetics and population analysis

MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies

D. Vuckovic[1,*], P. Gasparini[1,2], N. Soranzo[3,4] and V. Iotchkova[3,5,*]

[1]Department of Medical, Surgical and Health Sciences, University of Trieste, 34100 Trieste, Italy, [2]Medical Genetics, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", 34100 Trieste, Italy, [3]Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton CB10 1HH, [4]Department of Haematology, University of Cambridge, Cambridge CB2 0AH and [5]EMBL-EBI, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 19, 2015; revised on March 24, 2015; accepted on April 18, 2015

Abstract

Summary: As new methods for multivariate analysis of genome wide association studies become available, it is important to be able to combine results from different cohorts in a meta-analysis. The R package MultiMeta provides an implementation of the inverse-variance-based method for meta-analysis, generalized to an *n*-dimensional setting.
Availability and implementation: The R package MultiMeta can be downloaded from CRAN.
Contact: dragana.vuckovic@burlo.trieste.it; vi1@sanger.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have been a powerful tool for genetic discovery for almost a decade. Results have shed light on many different biological processes from lipid metabolism to blood composition as well as social and behavioral patterns (Hindorff *et al*., 2014, www.genome.gov/gwastudies.). A key to the success of GWAS was the ability to combine several studies in a meta-study, which allowed sufficiently large sample sizes for powered association studies even for variants of small phenotypic effect.

Although typically GWAS test one phenotype at a time, many biological features are better described by a combination of several variables. Thus, addressing multiple phenotypes can give great increase in power, by taking into account the underlying correlations between variables. Multivariate regression is a straightforward generalization of standard GWAS, where a linear multivariate mixed model can be fit and allows controlling for population stratification and relatedness (Korte *et al*., 2012; Zhou and Stephens, 2014). In particular, a new set of algorithms included in the GEMMA software (Zhou and Stephens, 2014) allows for a fast multivariate fitting and testing up to 10 phenotypes in large sample sizes. As for the univariate GWAS, it might be desirable in some settings to be able to combine multivariate results from different studies to further increase statistical power in genotype–phenotype associations.

Here, we describe a novel statistically efficient method to perform meta-analysis in a multivariate setting. It is an inverse-variance-based method that allows different weights for each cohort to take into account the accuracy of each effect estimate. The inverse-variance method has been successfully used for single-trait association testing (Sanna *et al*., 2008) and is suitable for *n*-dimensional generalization. Finally, it is implemented as part of the R package MultiMeta to benefit from flexible environment and open access, as well as extra plotting functions for results visualization.

2 Methods

The multivariate setting implies that results for each single-nucleotide polymorphism (SNP) include several effect sizes (also known as beta-coefficients, one for each trait), as well as related variance and covariance values, since beta coefficients can be correlated. Let p be the number of phenotypes analyzed and n the number of cohorts to include in the meta-analysis. For each cohort $i \in \{1, \ldots, n\}$ let β^i be the $px1$ vector of effect sizes and Σ^i be its pxp variance–covariance matrix.

Input data:

$$\beta^i = \begin{pmatrix} \beta_1^i \\ \vdots \\ \beta_p^i \end{pmatrix}; \ \Sigma^i = \begin{pmatrix} \sigma_{11}^i & \cdots & \sigma_{1p}^i \\ \vdots & \ddots & \vdots \\ \sigma_{p1}^i & \cdots & \sigma_{pp}^i \end{pmatrix}$$

To combine effect sizes, we have developed an inverse-variance-based approach. It is an *n*-dimensional generalization of the single trait meta-analysis, such as the one implemented in METAL software (Willer *et al.*, 2010). In particular, each vector β^i is weighted by the inverse of its variance–covariance matrix $(\Sigma^i)^{-1}$, then the final effect size B is computed as the weighted mean of all the beta coefficients.

$$B = \left[\sum_i \left(\Sigma^i\right)^{-1}\right]^{-1} \sum_i \left(\Sigma^i\right)^{-1} \beta^i$$

The variance of B is:

$$\mathrm{Var}(B) = \left[\sum_i \left(\Sigma^i\right)^{-1}\right]^{-1}$$

The resulting standardized beta follows a multivariate normal distribution.

$$S = \mathrm{Var}(B)^{-1/2} \ B \sim \mathbb{N}_p(0, \ I)$$

$$S = \left[\sum_i \left(\Sigma^i\right)^{-1}\right]^{-1/2} \sum_i \left(\Sigma^i\right)^{-1} \beta^i$$

Finally, significance of the multivariate association is tested against a chi-squared distribution with p degrees of freedom.

$$S^T S \sim \chi_p^2$$

A special case arises when the variance–covariance matrix of the effect sizes is singular for at least one of the cohorts under consideration. This case can typically occur when there is collinearity between phenotypes and needs to be considered separately because all calculations presented here rely on invertibility of all covariance matrices (which translates to non-zero variances in the univariate case). In general, multivariate modeling under such scenarios is probably best to be avoided; however, to get approximate meta-analysis estimates, a very small value can be added to the diagonal of the effect size covariance matrix so as to make it non-singular.

3 Results

The method was implemented as part of the R package *MultiMeta*, together with plotting functions useful for visual representation of the results, including Manhattan plot, quantile–quantile plot and an overview plot of effect sizes for a chosen SNP. The default options for the meta-analysis function *multi_meta* are set to work with GEMMA file format (multivariate analysis option). The plotting functions work with output files from the *multi_meta* function by default. However, both can be easily adapted to deal with different file formats by changing options, such as field separators, or by changing column names, as specified in the manual. Furthermore, the plotting functions can be run by passing files and objects in input. This choice is meant to provide more flexibility and avoid unnecessary opening of large files. Example datasets with only few SNPs are included in the package and can be analyzed as detailed in the instruction manual.

To test the software, we analyzed UK10K data (two cohorts, total sample size 3621) on six lipid-related traits: apolipoprotein A (ApoA), apolipoprotein B, high-density lipids (HDL), low-density lipids, triglycerides and total cholesterol (Supplementary Table S1). Results were compared with univariate GWAS meta-analysis on the same sample (Table 1). Seven known associations were confirmed. Interestingly, two of these (LIPC and PLTP) did not reach genome-wide significance in the univariate GWAS meta-analysis in our sample but did using the multi-trait approach, clearly showing a gain in power with the latter. Conversely, the multivariate analysis did not detect one locus on chromosome 11, suggesting that a single test is not necessarily the most powerful and that the two approaches should be complementary (Zhou and Stephens, 2014). Supplementary Figure S1 shows the plot obtained with *betas_plot* function for one of the significant SNPs.

Table 1. Summary of known loci reaching genome-wide levels of significance in comparisons of multivariate and univariate analysis

Chr	SNP[a]	Multi-trait *P* value	Single-trait *P* value	Single trait analysis	Known gene
1	rs602633	4.09E-09	1.74E-09	LDL	PSRC1
1	rs660240	2.97E-09	3.87E-10	ApoB	CELSR2
11	rs964184	0.003	6.81E-09	TG	ZPR1
15	rs1077835	2.06E-08	8.38E-05	HDL	LIPC
16	rs3764261	2.10E-40	4.98E-38	HDL	CETP
19	rs7412	1.01E-76	9.80E-68	ApoB	APOE
20	rs6065904	3.01E-12	0.002249	ApoA	PLTP

Chr, chromosome; single trait analysis, which trait in the univariate GWAS showed the strongest association for the SNP; LDL, low-density lipid; TG, triglycerides; ApoB, apolipoprotein B.

[a]Only previously reported SNPs are shown (www.genome.gov/gwastudies).

The meta-analysis runs on very low memory with default options (e.g. RAM < 250 Mb for two cohorts). Computation time is ~0.07 s/SNP for two cohorts and it grows linearly with the addition of other cohorts in input. As each SNP is analyzed singularly, overall computation time depends linearly on the total number of SNPs. By changing settings to increase the dimension of regions in which the genome is divided (see manual), it is possible to increase the performance, while allocating more memory.

The package is freely available on CRAN repository and can be run on any operating system.

4 Conclusion

Combining results from different cohorts is particularly important for GWAS, where large sample sizes are required to reliably detect alleles with small effects. The R package *MultiMeta* provides a flexible approach to meta-analyzing multivariate GWAS and easily visualizing results.

Acknowledgements

UK10K: This study makes use of data generated by the UK10K Consortium, derived from samples from the ALSPAC and TwinsUK datasets. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org.

Funding

Funding for UK10K was provided by the Wellcome Trust under award WT091310. This work was supported by the Wellcome Trust (Grant Codes WT098051 and WT091310 to N.S.), the NIHR BRC (to N.S.) and the EU FP7 (EPIGENESYS Grant Code 257082 and BLUEPRINT Grant Code HEALTH-F5-2011-282510 to N.S.).

Conflict of Interest: none declared.

References

Hindorff,L. *et al.* (2014) *A Catalog of Published Genome-Wide Association Studies*. http://www.genome.gov/GWAStudies/ [August 2014, date last accessed]

Korte,A. *et al*. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.*, **44**, 1066–1071.

Sanna,S. *et al*. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.*, **40**, 198-203.

Willer,C.J. *et al*. (2010) METAL: fast and efficient meta-analysis of genome-wide association scans. *Bioinformatics*, **26**, 2190–2191.

Zhou,X. and Stephens,M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.