Bioinformatics, 33(19), 2017, 3129–3130
doi: 10.1093/bioinformatics/btx395
Advance Access Publication Date: 15 June 2017
Applications Note

OXFORD

Genetics and population analysis

rqt: an R package for gene-level meta-analysis

Ilya Y. Zhbannikov*, Konstantin G. Arbeev and Anatoliy I. Yashin

Biodemography of Aging Research Unit (BARU), Social Science Research Institute, Duke University, Durham, NC 27708, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on January 19, 2017; revised on May 22, 2017; editorial decision on June 12, 2017; accepted on June 13, 2017

Abstract

Motivation: Despite recent advances of modern GWAS methods, it is still remains an important problem of addressing calculation an effect size and corresponding *p*-value for the whole gene rather than for single variant.
Results: We developed an R package *rqt*, which offers gene-level GWAS meta-analysis. The package can be easily included into bioinformatics pipeline or used stand-alone. We applied this tool to the analysis of Alzheimer's disease data from three datasets CHS, FHS and LOADFS. Test results from meta-analysis of three Alzheimer studies show its applicability for association testing.
Availability and implementation: The package *rqt* is freely available under the following link: https://github.com/izhbannikov/rqt.
Contact: ilya.zhbannikov@duke.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in genome-wide analyses of complex traits allowed for detecting a number of strong associations between genetic factors and chronic diseases (Kathiresan *et al.*, 2009; Strawbridge *et al.*, 2011). However, many other associations detected in such studies are weak and did not reach genome wide levels of statistical significance. These results are in concert with the Fisher's conjecture that genetic variability of complex traits is the result of integration of the influence of many genetic factors each having small effect on the trait (Fisher, 1999). This situation generates an idea that evaluation of components of genetic influence on complex traits that are integrated by some parts of biological mechanisms may improve strength of the genetic estimates. Attempts to realize this idea resulted in a number of statistical methods focused on gene-level analyses of genetic data. In such analyses, information about detected genetic variants that belong to a particular gene is used to construct a score variable that integrates associations of these variants with a disease within a gene. Several methods were developed for such integration of common- and rare variants. This includes SKAT (Wu *et al.*, 2011), KBAC (Liu *et al.*, 2010), WSS (Madsen *et al.*, 2009) and others. A set of methods and tools that perform gene-level meta-analysis was also proposed, e.g. MetaSKAT (Lee *et al.*, 2013), seqMeta (https://CRAN.R-project.org/package=seqMeta). The procedures of integrating genetic signals used in these methods are based on different ideas and result in different estimates of integrated associations of selected genes with the traits of interest. The methods that produce stronger genetic associations with phenotypic traits are usually considered as more preferable.

Recently, the QTests (Lee *et al.*, 2016) for rare variants have been proposed. In these regression-based tests (QTest1-3), a score for the whole gene is built by using the pooled effect size obtained as a weighted sum of corresponding effect sizes from fitting a multivariate regression on all detected variants in a gene. However, it is important to take possible presence of linkage disequilibrium (LD) into account. LD is directly related to multicollinearity, which can significantly impact analysis by increasing the variance of the coefficient estimates (thereby making them very sensitive to small changes in the model); also the estimates become unstable and can switch signs. In addition, multicollinearity has negative effect on power (Yoo *et al.*, 2014).

In this note we propose RQTests, which are modified QTests. RQTests directly address the problem of possible LD between variants . We also present a corresponding software tool—an R package *rqt*, which performs gene-level and meta-analyses taking into account rare genetic variants. *rqt* is available for download from the following link: https://github.com/izhbannikov/rqt.

2 Materials and methods

The workflow of gene-level meta-analysis consists of the following steps: (i) reducing the number of predictors (a.k.a. 'data preprocessing') to exclude multicollinearity. This is the primary contribution of our work; (ii) then the regression model is fitted on the reduced dataset to obtain corresponding regression coefficients; (iii) these coefficients are used to construct statistics representing a gene-level effect. *P*-values are then calculated using this statistics with asymptotic approximation or permutation procedure; (iv) combining gene-level *p*-values calculated from each study. Below we describe these steps in details:

- In order to alleviate effects of correlation (LD-effects) between variants we employ a set of methods to preprocess the data first: PCA (principal component analysis), PLS (partial least square), LASSO (Tibshirani, 1996) and ridge regressions. By default, we use PCA and the number of principal components is used to capture 75% of explained variance. We should note that in PLS regression both the original predictors and response are decomposed into latent structures therefore the final *p*-values and pooled regression coefficient (see below) are estimated with respect to those *new* variables. The user can avoid this step (and, thereby, choose QTests) by supplying 'none' in the `method` parameter from funtction `geneTest(...)` of *rqt*.
- Pooling regression coefficients and calculating statistics and gene-level *p*-value is performed according to the method proposed in (Lee *et al.*, 2016). The pooled effect size $\widehat{\beta}_{Pooled}$ is a weighted sum of coefficients: $\widehat{\beta}_{Pooled} = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{W}\widehat{\boldsymbol{\beta}}$, where $\boldsymbol{\alpha} = (\alpha_k)_{m\times 1}$, $\alpha_k = \frac{1/var(\hat{\beta}_k)}{\sum_{k=1}^{m}(1/var(\hat{\beta}_k))}$ - an inverse variance vector for the estimates of $\widehat{\beta}_k$ from the multivariate regression. $\boldsymbol{W}$ is a diagonal weight matrix that contains weighs for *j*-th variant: $w_j = Beta(MAF_j, 1, 25)$ where MAF_j is a corresponding minor allele frequency; *m* is the number of variants in a gene. Then the corresponding statistics Q_1 for the QTest1 is calculated as follows:

$$Q_1 = (\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{\alpha})^{-1}\widehat{\beta}_{Pooled}^{\ 2} \sim \chi^{2}_{-1},$$

where $V = \mathrm{var}(\widehat{\boldsymbol{\beta}})$. Other test statistics (Q_2 for QTest2 and Q_3 for QTest3) are described in (Lee *et al.*, 2016). In our package *rqt* we implemented all of them. Since RQTests are implemented on top of QTests, RQTests1-3 take into account rare variants as well.

- Calculating combined *p*-value for a gene from several studies is performed with one of the available combing probability methods (refer to the User Manual.)

3 Results

We developed an R package, *rqt*, where we implemented and improved QTests (rQTests1-3.) We also applied this methodology to a set of genes that are potentially involved in the development of Alzheimer's disease (AD), obtained from a literature search. We conducted a meta-analysis of the following studies: CHS (Cardiovascular Health Study, dbGaP accession: phs000287.v5.p1), FHS (Framingham Heart Study, dbGaP accession: phs000007.v22.p8) and LOADFS (Late Onset Alzheimer's Disease Family Study, dbGaP accession: phs000168.v2.p2), both females and males, in order to evaluate the possible associations between these genes and AD. Supplementary Table S1A from Supplementary Materials presents a description of the datasets used. Table 1 shows results for genes showed (*P*-value $< 10^{-2}$) associations to AD. These results are concordant to those previously found (Corder *et al.*, 1993; Linnertz *et al.*, 2014).

Table 1. Results of gene-level and gene-level meta-analysis: top genes with *P*-value less than 10^{-2} sorted by *P*-value for rQTest3

Gene	*P*-value			
	Meta	CHS	FHS	LOADFS
PVRL2	3.000E−09	1.030E−09	3.629E−01	1.00E−09
TOMM40	3.000E−09	3,300E−08	2.600E−01	1.00E−09
APOC1	9.030E−06	3,010E−06	5.292E−02	6.34E−05
PPP1R3B	4.092E−04	1,364E−04	9.995E−01	1.469E−02
APOE	3.512E−03	2.048E−01	3.704E−01	1.172E−03

Detailed analysis results by study are shown in Supplementary Materials, Tables 3C–6C.

We also performed power simulation and type 1-error tests for *rqt* (RQTest1-3), QTest1-3, SKAT and SKAT-O (see Supplementary materials for simulation setup), for dichotomous and continuous phenotypes, assuming presence of LD between variants. Sample size was 3,000 and 50 SNPs. Percentage of causal SNPs was 10% and 25% in each test. Simulation methodologies and results are shown in Supplementary Figures S1C–S14C and Tables S7C–S13C from Supplementary. RQTests offer lowest variance inflation factor (VIF), type 1 error rate and highest power for the PCA preprocessing method, see Supplementary Figures S1C, S2C, and Tables S7C and S8C from Supplementary materials. Note: according to the QQ plots under the null hypothesis for the *PLS* method, the *p*-values have non-uniform distribution and therefore the *PLS* method should be used with care in *rqt*. Additional investigations are needed to address this issue.

Funding

This work was supported by the National Institute on Aging of the National Institutes of Health (NIA/NIH) under Award Numbers P01AG043352, R01AG046860 and P30AG034424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIA/NIH. Acknowledgements on data used are given in Supplementary materials, Section D.

Conflict of Interest: none declared.

References

Corder,E. *et al.* (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923. p

Fisher,R. (1999) *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford, UK: Oxford University Press.

Kathiresan,S. *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56–65.

Lee,J. *et al.* (2016) Gene-set association tests for next-generation sequencing data. *Bioinformatics*, **32**, i611–i619.

Lee,S. *et al.* (2013) General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 42–53.

Linnertz,C. *et al.* (2014) The cis-regulatory effect of an Alzheimer's diseaseas-soc ated poly-T locus on expression of TOMM40 and apolipoprotein E genes. *Alzheimer's Dement. J. Alzheimer's Assoc.*, **10**, 541–551.

Liu,D. *et al.* (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, 1–14.

Madsen,B. *et al.* (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, 1–14.

Strawbridge,R. *et al.* (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*, **60**, 2624–2634.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

Wu,M. *et al.* (2011) Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.*, **89**, 82–93.

Yoo,W. *et al.* (2014) A study of effects of MultiCollinearity in the multivariable analysis. *Int. J. Appl. Sci. Technol.*, **4**, 9–19.