Bioinformatics, 32(4), 2016, 624–626
doi: 10.1093/bioinformatics/btv609
Advance Access Publication Date: 27 October 2015
Applications Note

Genetics and population analysis

RVFam: an R package for rare variant association analysis with family data

Ming-Huei Chen[1,3] and Qiong Yang[2,3,*]

[1]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA, [2]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA and [3]Framingham Heart Study, Population Sciences Branch, Division of Intramural Research, National Heart Lung and Blood Institute, National Institutes of Health, Framingham, MA 01702, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on March 11, 2015; revised on September 16, 2015; accepted on October 16, 2015

Abstract

Summary: Family-based designs offer unique advantage for identifying rare risk variants in genetic association studies. There are existing tools for analyzing rare variants in families but lacking components to handle binary traits properly and survival traits. In this report, we introduce an R software package RVFam (Rare Variant association analysis with Family data) designed to analyze continuous, binary and survival traits against rare and common sequencing variants in genome-wide association studies (GWAS) involving family data. Single and multiple variant association tests were implemented while accounting for arbitrary family structures. Extensive simulation studies were performed to evaluate all the approaches implemented in RVFam.

Availability and Implementation: http://cran.r-project.org/web/packages/RVFam/

Contact: qyang@bu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It was hypothesized that rare variants poorly represented in existing GWAS may underlie the unexplained heritability by common variants identified to date. Emerging exome sequencing and whole genome-sequencing studies are designed to capture rare variants in human genomes. The regularity condition assumed in many existing GWAS tools for common variants may become invalid when applied to rare variants (Chen *et al.*, 2011, Chen *et al.*, 2014). Family-based designs are advantageous for identifying rare risk variants such as private or pedigree specific mutations that may only be enriched with family samples. Ignoring family structure could introduce bias in the results and selecting unrelated sample from family data reduces power. There are existing tools that can handle rare variant analysis with family samples, such as rareMetalWorker (http://genome.sph.umich.edu/wiki/RAREMETALWORKER), seqMeta (https://cran.r-project.org/web/packages/seqMeta), and EPACTS (http://genome.sph.umich.edu/wiki/EPACTS). The three packages report test based on score statistic from linear mixed effects model (LME) for continuous traits measured on family samples. For binary traits, seqMeta can only analyze unrelated sample, and RareMetalWorker and EPACTS use LME and treat binary traits as continuous, so the effect estimate is not interpretable. None of the package can handle survival traits. In this report, we introduce RVFam that is designed to provide tools for family samples for all three types of traits: continuous, binary and survival. RVFam can perform single variant and multiple variant pooled analysis for a single cohort as well as producing score statistics formatted for meta-analysis in seqMeta. We performed extensive simulation studies to assess the approaches implemented in RVFam.

2 Methods

2.1 Single variant analysis

For continuous traits, we use LME with a fixed effect for genotype score, and with person specific random intercepts that are correlated only within family according to relationship coefficients

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

(Supplementary Methods). This model is implemented in RVFam by calling lmekin() in R coxme package (version 2.2-3). Different from aforementioned packages that report test based on score statistic, Wald test result is reported in RVFam. For binary traits, we implemented generalized linear mixed effects model (GLMM) with a logistic link and with same fixed and random effects as LME except an exchangeable within family correlation structure for random effects. This model is implemented in RVFam by calling glmer() in R lme4 package (version 1.1-7). For survival traits, Cox-proportional hazard with shared frailty (random effect) in each family is implemented by calling coxph() in R survival package (version 2.37-7). Likelihood ratio (LR) test *P*-values are reported for binary and survival traits as inflation was observed for Wald test in our simulations (Supplementary Table S2A, S3A). The methods and tests for all three types of traits (Supplementary Methods) were chosen based on statistical properties as well as results from our extensive simulation studies.

We also provide interface with seqMeta so that RVFam results can be directly meta-analyzed in seqMeta. The required R object by seqMeta for meta-analysis contains the score statistic and its variance for each SNP that was converted from our Wald or LR test (Supplementary Methods).

2.2 Gene-based analysis

For gene-based analysis, we implemented two burden tests ((Li and Leal, 2008) denoted by *T*, and (Madsen and Browning, 2009) denoted by MB) that are powerful when the directions of association are the same across multiple variants, and one direction insensitive test: the sum of squares (SSQ) test (Pan, 2009). For burden tests, a super variant is created by summing the weighted genotype score of selected SNPs within a gene region (weight = 1 for *T* and $(q(1-q))^{-1}$ for MB, where q is minor allele frequency (MAF)). The super variant is analyzed using the same method as in single variant analysis. For SSQ, *Z*-statistics or signed LR statistics from single variant analysis are squared and summed up for selected SNPs within a gene or region. Under null hypothesis that none of the SNPs are associated with the outcome, SSQ $\sim \chi^2(n)$, where *n* is the degrees of freedom depending on the covariance matrix among the statistics (Supplementary Methods). We have previously shown through empirical studies that the SSQ had similar power as SKAT for unrelated sample (Wang *et al.*, 2012).

3 Input and output

Input to this package are SNP genotype data coded as 0, 1, and 2 for number of copies of coded allele, phenotype and covariate data, and pedigree data. It also requires gene annotation for each SNP for forming pooled multi-variant tests, a comma delimited file with MAF (based on all genotyped sample), and an RData containing genotype correlation matrix for the defined genes or regions. The output from RVFam includes: (i) a text file containing single variant test results including beta, se, and *P*-value; (ii) text files of gene-based test results; (iii) an RData containing score statistics that can be directly used by seqMeta for meta-analyses. See RVFam Document in Supplementary information for details of input and output.

4 Simulation studies

We conducted extensive simulation studies to evaluate the validity and power of the tests implemented in RVFam. The simulations used the real genotypes of 225 160 nonsynonymous, stop-altering, and splice variants captured on the Illumina HumanExome BeadChip, and simulated phenotypes of 3880 Framingham Heart Study (FHS) Offspring samples from 1147 families. Description of the genotyping was reported previously (Peloso *et al.*, 2014). In summary, about 90% of the variants have a MAF 0.05 or less, and 50% have MAF lower than 0.0001. We used SOLAR (Almasy and Blangero, 1998) to simulate phenotypes conditional on the observed family structures in FHS sample. Survival traits were simulated to follow a Weibull distribution with normal distributed random effects incorporated in the scale parameter. Binary traits were simulated based on an additive threshold model. To evaluate the validity of the methods, 100 replicates of phenotypes were generated independent of all genetic variants, with a polygenic heritability 0.25 for the random effects. To evaluate power, we assign various effects to 5 SNPs (pairwise $R^2 < 0.01$, MAF 0.0003, 0.001, 0.004, 0.02, 0.25, respectively) in the ABO gene that contains a total of 36 SNPs (Supplementary Table S1-B). We took the event variable from survival traits as our binary traits to evaluate power. MAF ranges of (0, 0.01) and (0, 0.05) were used for gene-based analysis.

4.1 Simulation results

The type I error and power for RVFam analysis of continuous traits are presented in Supplementary Table S1-A,B,C. We found that the single variant and gene-based tests (Supplementary Table S1-A) yielded valid genome-wide type I error rates (<0.05) as well as SNP-wise and gene-wise type I error rates (2.22×10^{-7} and 2.85×10^{-6}, respectively, by Bonferroni correction). The power simulations (Table S1-B,C) show that gene-based tests can achieve higher power than single variant test, and SSQ test is more powerful than T and MB especially when QTL of opposite effects are included (SSQ5).

The type I error and power for RVFam analysis of binary traits are presented in Supplementary Table S2-A, B. RVFam had valid type I error rates when prevalence was 5% or higher. Inflation was reduced significantly by applying a minor allele count filter among cases when prevalence is 0.01 (Supplementary Table S2-A). Power results show similar pattern as described in continuous traits.

The type I error and power for RVFam analysis of survival traits are presented in Supplementary Table S3-A, B. We observed valid type I error rates for RVFam single SNP analysis and T but slight inflation for MB and SSQ tests in some cases. To explore this further, we calculated type I error by MAF groups and found inflation only in SNPs with MAF < 0.001. This explains why MB is inflated because it weights each SNP by its inverse MAF and highlights extremely low MAF SNPs. By aggregating single SNP results, inflation also occurs to SSQ for genes with multiple extremely low MAF SNPs. We propose a remedy to reduce inflation in low MAF single SNP results and SSQ that divides the single SNP chi-square statistics by the genomic control parameter for rare SNPs (e.g. minor allele counts < 5) and then re-computes SSQ. Power results (S3-C) show similar pattern as described in continuous and binary traits.

It took RVFam 46, 64 and 19 min for continuous, binary and survival traits, respectively, to complete all single variant and gene-based tests for chromosome 21 containing 2671 SNPs and 254 genes for a sample of 400 three-generation pedigrees each with 10 members on a linux cluster with 4 16-core 2.3 GHz AMD Opteron 6276 processors and 256 GB of RAM running CentOS 6 with Sun Grid Engine.

Acknowledgements

The authors thank Arend Voorman for coming up with the conversion of our single variant results to the score statistics for seqMeta and Katia Oleinik for improving the efficiency of our scripts. This research was conducted in part using data and resources from the Framingham Heart Study and Boston University School of Medicine.

Funding

This work was partially supported by the National Institutes of Health (NIH)/ National Heart, Lung, and Blood Institute (NHLBI)'s Framingham Heart Study contract (N01 HC25195 and N01 HHSN-268201500001). A portion of this research utilized the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center.

Conflict of Interest: none declared.

References

Almasy,L. and Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.

Chen,H. *et al.* (2014) Sequence kernel association test for survival traits. *Genet. Epidemiol.*, **38**, 191–197.

Chen,M.H. *et al.* (2011) A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genet. Epidemiol.*, **35**, 650–657.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Pan,W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.*, **33**, 497–507.

Peloso,G.M. *et al.* (2014) Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56 000 whites and blacks. *Am. J. Hum. Genet.*, **94**, 223–232.

Wang,Y. *et al.* (2012) Joint rare variant association test of the average and individual effects for sequencing studies. *PLoS One*, **7**, e32485.