Using RVFam package to conduct

rare variant analysis with family data

Ming-Huei Chen and Qiong Yang

Departments of Neurology and Biostatistics

Boston University

Contact: [qyang@bu.edu](mailto:qyang@bu.edu)

**Table of Contents**

[Overview 1](#_Toc241486280)

[Methods 1](#_Toc241486281)

[Required Files 2](#_Toc241486282)

[Examples 4](#_Toc241486283)

[Output](#_Toc241486285) 5

[References](#_Toc241486285) 7

## Overview

This package (RVFam) was designed to analyze rare variants against a continuous, or dichotomous, or survival phenotype measured on subjects from families for genetic association. For each major function call (e.g. glmm.binped, lme.ped, coxph.ped), single SNP analysis and three gene-based tests (two burden tests and one sum of squares test) are applied to analyze rare variants on a chromosome.

## Methods

Linear mixed effects (LME) model is used to analyze continuous traits. In LME modeling, a random intercept, person specific random effects correlated according to degree of relatedness (i.e. kinship coefficient) within a family, is used to account for within family correlation. Generalized linear mixed effects model (GLMM) with logistic link and a normal random intercept for each cluster/pedigree/family is used to analyze dichotomous traits. Cox proportional hazards (COXPH) regression model with shared frailty (random effects) is used to analyze survival traits. Our package calls lmekin() function from coxme package, glmer() function from lme4 package, and coxph() function from survival package, for LME, GLMM and COXPH.

In addition to regular single SNP analysis, RVFam package also implements three gene-based tests, including two burden tests (weight=1 for Li & Leal 2008; weight=1/(MAF)/(1-MAF) for Madsen & Browning 2009), one sum of squares (SSQ) test (Wei 2009). For single SNP analysis and burden tests, in LME, Wald Z statistic is reported and p-value is computed based on Wald chi-square statistic (Z2), while in GLMM and COXPH, signed LRT statistic is reported and p-value is computed based on LRT. A remark column is given in single SNP analysis and burden test output files to provide warning or additional information for the analysis. For GLMM, a remark of converging issue (returned by lme4 package) indicates that estimation procedure did not converge - user should pay attention to it and may consider excluding results with converging issue particularly when beta/se estimates are larger than expected. An RData output that can be used with seqMeta package for meta-analysis is also generated.

## Required Files

Before performing analyses with this package, following files have to be created.

1. **Pedigree file**: A file containing all the families is required. The column names should be exactly the same (case sensitive) as in following example based on **comma delimited format**. Missing father (fa) or mother (mo) ids should be 0. Individuals who are unrelated to anyone can be included as family of size 1.

famid,id,fa,mo,sex

1,10,0,0,1

1,11,0,0,2

1,12,10,11,1

1,13,10,11,1

3,32,0,0,1

3,33,0,0,2

3,334,32,33,1

3,335,32,33,2

10,50,0,0,1

11,60,0,0,2

1. **Phenotype and genotype files**

**Phenotype file** contains unique individual id, phenotype and covariates. The header should contain “id”, followed by other variable names. Use empty space for missing values. **Dichotomous phenotype must be coded as 0, 1 with 1 being affected**. Covariates values must be coded numerically (dichotomous covariate can have any two numeric values). Following is an example of the phenotype file based on **comma delimited format**:

id,phen1,phen2,covar1,covar2

10,100,1,1,0.2

112,,0,1,0.3

312,130,1,2,0.4

513,125,0,,0.5

**Genotype file** contains unique individual id and genotype data. The header should contain “id”, followed by SNP names. G**enotypes should be coded as 0, 1, 2 representing the copies of the coded allele**. Use NA for missing genotypes. SNP names should not contain special characters such as “-“,”/”, etc. But “.” and “\_” are allowed. For example (based on **comma delimited format**):

id,SNP.1,SNP\_2

10,0,1

11,,

12,1,2

13,2,0

1. **SNP info R data:** An RData file that contains at least SNP name (“Name”), chromosome number “Chr”, and the column name for aggregating/grouping SNPs (default is “SKATgene”).
2. **MAF file:** A file in comma delimited format that contains SNP name (“Name”) and minor allele frequency (“maf”).
3. **SNP correlation matrix RData:** An RData file that contains SNP correlation matrices, where SNPs are aggregated by “aggregateBy”.

## Examples

Here are example function calls for analyzing a single phenotype against all exome chip SNPs on chromosome 21. Suppose “phenotype.csv”, “ped.csv”, “chr21geno.csv”, “MAFchr21.csv”, “chr21snpinfo.RData”, “chr21.SNPs.in.gene.Cor.RData”,are directory paths to **comma delimited** phenotype, genotype, MAF and pedigree files, and SNP info and SNP correlation matrix RData, respectively; “simqt” and “aff” are the name of the continuous and binary phenotype to be analyzed, respectively. In LME, kinship coefficient matrix is computed based on input pedigree file. In addition to “aff”, a time variable “time” is used for analyzing survival traits.

RVFam includes SOLAR-simulated (Almasy L. & Blangero J. 1998) genotype (chr21geno.csv) and phenotype (phenotype.csv) data based on a fake sample that contains 400 pedigrees (ped.csv) each with 3 generations and 10 members of the same pedigree structure. The simulated genotype data were used to estimate MAF and generate MAF file (MAFchr21.csv), and to compute SNP correlation matrix RData (chr21.SNPs.in.gene.Cor.RData). The SNP info RData (chr21sninfo.RData) is also included. The csv (comma delimited) files are gzip’ed and can be accessed by the following command after library(RVFam). One needs to first save the RData and csv files (gunzip first) to try the examples.

> system.file("extdata", "ped.csv.gz", package="RVFam")

[1] "/usr2/faculty/mhchen/R/x86\_64-unknown-linux-gnu-library/3.1/RVFam/extdata/ped.csv.gz"

**LME:**

library(RVFam)

lme.ped(phenfile="phenotype.csv",phen="simqt",covars="sex",mafRange=c(0,0.05),chr=21,genfile="chr21geno.csv",pedfile='ped.csv',snpinfoRdata='chr21snpinfo.RData',sep.ped=',',sep.phe=',',sep.gen=',',aggregateBy='SKATgene',maf.file='MAFchr21.csv',snp.cor='chr21.SNPs.in.gene.Cor.RData',ssq.beta.wts=c(1,25))

**GLMM:**

library(RVFam)

glmm.binped(phenfile="phenotype.csv",phen="aff",covars=NULL,mafRange=c(0,0.05),chr=21,genfile="chr21geno.csv",pedfile='ped.csv',snpinfoRdata='chr21snpinfo.RData',sep.ped=',',sep.phe=',',sep.gen=',',aggregateBy='SKATgene',maf.file='MAFchr21.csv',snp.cor='chr21.SNPs.in.gene.Cor.RData',ssq.beta.wts=c(1,25))

**COXPH:**

library(RVFam)

coxph.ped(phenfile="phenotype.csv",phen="aff",covars="sex",mafRange=c(0,0.05),chr=21,genfile="chr21geno.csv", pedfile='ped.csv',snpinfoRdata='chr21snpinfo.RData',sep.ped=',',sep.phe=',',sep.gen=',',time="time", aggregateBy='SKATgene',maf.file='MAFchr21.csv',snp.cor='chr21.SNPs.in.gene.Cor.RData',ssq.beta.wts=c(1,25))

Important: These functions are designed to analyze a single phenotype against all the SNP genotypes on a chromosome in a genotype file in a single call. To analyze multiple phenotypes, multiple calls of the functions are needed.

## Output

**Output information**: Output from a major function call is saved to the files specified by trait name, chromosome number, type of tests (singleSNP, T, MB and SSQ), and mafRange in each major function (lme.ped, glmm.binped and coxph.ped). An RData is also generated and can be used with seqMeta package for meta-analysis. All three types of analyses (LME, GLMM and COXPH) have the same format of output files. Table 1 describes the output columns for single SNP analyses. Table 2 describes the output columns for gene-based burden tests (T and MB). Table 3 describes the output columns for the SSQ test.

**Table 1: Output columns from single SNP analysis (Genotype should be coded as 0, 1, 2 representing the copies of the coded allele)**

|  |  |
| --- | --- |
| **Column** | **Description** |
| **gene** | gene name |
| **Name** | SNP name |
| **maf** | MAF estimate based on genotyped sample and reported in MAF file |
| **ntotal** | number of individuals with genotype, phenotype and covariates |
| **nmiss** | number of individuals with missing genotype among ntotal |
| **maf\_ntotal** | MAF estimate based on analysis sample |
| **beta** | beta coefficient per 1 copy increment of coded allele |
| **se** | Standard error of beta |
| **Z** | Z statistic, GLMM and COXPH report signed LRT; LME reports Wald Z |
| **remark** | warning or additional information for the analysis |
| **p** | p-value, GLMM and COXPH use (sign) LRT; LME uses Wald chi-square test |
| **MAC** | Minor allele count |
| **n0** | number of subjects with non-missing phenotype and genotype 0 |
| **n1** | number of subjects with non-missing phenotype and genotype 1 |
| **n2** | number of subjects with non-missing phenotype and genotype 2 |

**Table 2. Output columns from Burden tests (T and MB)**

|  |  |
| --- | --- |
| **Column** | **Description** |
| **gene** | gene name |
| **beta** | beta coefficient of aggregated super variant |
| **se** | Standard error of beta |
| **Z** | Z statistic, GLMM and COXPH report signed LRT; LME reports Wald Z |
| **cmafTotal** | sum of maf\_ntotal of SNPs in a gene |
| **cmafUsed** | sum of maf\_ntotal of selected SNPs in a gene |
| **nsnpsTotal** | Total number of SNPs in a gene |
| **nsnpsUsed** | Total number of selected SNPs in a gene |
| **nmiss** | Sum of nmiss’s of selected SNPs in a gene |
| **remark** | warning or additional information for the analysis |
| **p** | p-value, GLMM and COXPH use (sign) LRT; LME uses Wald chi-square test |

**Table 3: Output columns from SSQ test**

|  |  |
| --- | --- |
| **Column** | **Description** |
| **gene** | gene name |
| **SSQ** | Sum of squares statistic |
| **cmafTotal** | sum of maf\_ntotal of SNPs in a gene |
| **cmafUsed** | sum of maf\_ntotal of selected SNPs in a gene |
| **nsnpsTotal** | Total number of SNPs in a gene |
| **nsnpsUsed** | Total number of selected SNPs in a gene |
| **nmiss** | Sum of nmiss’s of selected SNPs in a gene |
| **df** | Degrees of freedom of SSQ |
| **p** | p-value, GLMM and COXPH use (sign) LRT; LME uses Wald chi-square test |

## References

Li, B. and Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet*, 2008; **83(3)**, 311-321.

Madsen, B. E. and Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet*, 2009; **5(2)** e1000384.

Wei P. Asymptotic Tests of Association with Multiple SNPs in Linkage Disequilibrium.

*Genet Epidemiol*, 2009; **33(6)**, 497-507.

Almasy L. and Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 1998; **62(5)**:1198-211.