RIA: Regional IBD Analysis

Contents

1 Introduction

The program RIA is a C++ implementation of the method described in Eu-Ahsunthornwattana et al. (2015) and uses calls to programs PLINK and KING version 2.2.9.

1.1 Overview

The program RIA performs the following steps calling external PLINK and KING where required:

1. The first stage of RIA is the estimation of the IBD sharing probabilities between affected family members using a selection of SNPs across all available SNPs in the data, these probabilities form the *priors*. The program PLINK Purcell et al. (2007) is used to prune the SNPs to give a representative selection of SNPs using only cases which are then used to estimate the IBD sharing probabilities using KING (version 2.2.9) Manichaikul et al. (2010). See section 1.2 on how the IBD sharing probabilities are estimated.

2. The next stage is to step across the genome, set to a default of 50 SNPs per step, and form a SNP window of around 500 to 2000 SNPs and use these SNPs to estimate the IBD sharing probabilities between affected family members using KING (version 2.2.9). These probabilities give the *posteriors*. See section 1.2 on how the IBD sharing probabilities are estimated.

3. The next step is to perform a non-parametric linkage analysis comparing the prior and posterior IBD sharing probabilities using the method as described in Cordell et al. (2000) with minor modifications as described in Eu-Ahsunthornwattana et al. (2015). This produces a LOD score for each analysed SNP region together with parameter estimates for the (scaled) additive and dominance variances.

For full details of the methodology of RIA please see the accompanying paper Eu-Ahsunthornwattana et al. (2015).

1.2 Estimation of IBD sharing probabilities

KING version 2.2.9 is used with option `--homog` to estimate the IBD sharing probabilities. Let *IBD2*, *IBD1* and *IBD0* be the probabilities that two individuals share 2, 1 or 0 alleles IBD respectively and K the kinship coefficient, then *IBD2*, *IBD1* and *IBD0* are estimated for each SNP as follows:

1. KING version 2.2.9 is ran with the `--homog` option which returns estimates of K and *IBD0*.

2. K and *IBD0* are truncated to plausible values between 0 to 0.5 and between 0 to 1 respectively if necessary.

3. *IBD2* is estimated as *4K - (1 - IBD0)*.

4. *IBD2* is truncated to a plausible value between 0 to 1 if necessary.

5. *IBD1* is estimated as *1 - IBD0 - IBD2*.

6. *IBD1* is truncated to a plausible value between 0 to 1 if necessary.

7. For the posterior probabilities only: if any of the priors for *IBD2*, *IBD1* or *IBD0* are equal to zero then the corresponding posterior values are also set to zero. If this results in all IBD probabilities being equal to zero then an error is reported.

8. *IBD2*, *IBD1* and *IBD0* are scaled so that *IBD2 + IBD1 + IBD0 = 1* if necessary.

2 Installation

Download an executable file from the home page for your system and off you go, or do the following.

1. Download the code from the home page.

2. Compile it by typing something like the following:

```
g++ -O3 src/*.cpp -o ria
```

3. It is also necessary to having working copies of the programs PLINK and KING version 2.2.9 installed. **Note:** It is very important to download KING version 2.2.9 and not any later versions as we need to use the `--homog` option which is no longer available in later versions.

4. Start analysing your data with RIA!

5. If you are using Windows then you will need to uncomment the line below in the `main.h` file:

```
#define USING_WINDOWS
```

3 Using RIA

The program RIA takes a PLINK binary file as input (.bed/.bim/.fam) and produces a results file of the analysis. Basic usage of the program is given by typing:

```
./ria -i mydata.bed
```

The most likely options that will need to be used are how to run the programs PLINK and KING:

```
./ria -plink /home/me/my-programs/plink/plink -king /home/me/my-programs/king
    /king -i mydata.bed
```

Typing ria with no options will output usage details:

```
RIA: Regional IBD Analysis, v1.1
----------------------------------------------------------------
Copyright 2015-present Richard Howey, GNU General Public License, v3
Research Software Engineering, Newcastle University

Usage:
  ./ria [options] -i pedigree.bed
 or ./ria -pf parameterfile [pedigree.bed]

Options:
  -window-size-cm n    -- set window size to n centimorgans
  -window-min-snps m   -- set required minimium number of SNPs in a window to m
  -step-size s         -- step size of windows, s
  -start-snp a         -- start analysis from SNP number a
  -end-snp b           -- end analysis at SNP number b
  -start-snp-name x    -- start analysis from SNP name x
  -end-snp-name y      -- end analysis at SNP name y
  -job d t             -- job number d of t
  -i file.bed          -- input binary pedigree file, file.bed
  -o results.dat       -- output results file, results.dat
  -i-prior file        -- input prior IBDs, file
  -o-prior file        -- output prior IBDs, file
  -prior-only          -- calculate prior IBDs only
  -plink command       -- command used to run PLINK
  -plink-options "ops" -- PLINK pruning options used to calculate the prior
  -king command        -- command used to run KING
  -log results.log     -- log filename, results.log
  -ndv                 -- no dominance variance
  -so                  -- suppress output to screen

Default Options in Effect:
  -window-size-cM 15
  -window-min-snps 100
  -step-size 50
  -plink plink
  -king king
  -o riaResults.dat
```

See section 4 for an example of how to use RIA to analyse some data.

3.1 Parameter file

A parameter file, .pf, may be used with RIA instead of writing all of the options on the command line. To use a parameter file simply type:

```
./ria -pf myparameters.pf
```

The parameter file should be a text file with one option written on each line. For example, to perform the analysis above the file myparameters.pf would be as follows:

```
-plink /home/me/my-programs/plink/plink
-king /home/me/my-programs/king/king mydata.bed
-window-size-cm 15
-step-size 50
-i mydata.bed
-o myResults.dat
```

It is also possible to add comments to the file provided that the "-" character is not used, and to comment out any options by placing another character in front of any "-". For example, the above parameter file could be edited as follows:

```
# Command used to run PLINK
-plink /home/me/my-programs/plink/plink

# Command used to run KING
-king /home/me/my-programs/king/king

# SNP window size
-window-size-cm 15

# Number of SNPs to move to the next SNP window
-step-size 50

# The all important data to analyse
-i mydata.bed

# My lovely analysis results
-o myResults.dat

# I might try this later
#-i myOtherData.bed
```

3.2 Temporary Files

RIA requires the use of two other programs, PLINK and KING, and in order to use these programs several temporary files must be created. All of these files begin with

"tempRIA" and, before the file extension, end with the process ID of the current RIA job. They will typically appear similar to the following:

```
tempRIA-priors1-57703.prune.in
tempRIA-priors2-57703.bed
tempRIA-priors2-57703.bim
tempRIA-priors2-57703.fam
tempRIA-priors2-57703.kin
tempRIA-post-57703.fam
tempRIA-posterior-57703.bed
tempRIA-posterior-57703.bim
tempRIA-posterior-57703.fam
tempRIA-posterior-57703.kin
```

The use of the process ID ensures that several RIA jobs may be ran in the same location without any issues of interference. All of these temporary files should be deleted by RIA when they are no longer needed, however if RIA is forced to unexpectedly stop for some reason then these files may still be lying around. In which case, they should be carefully deleted if necessary.

4   RIA Example

This section runs through an example analysis using RIA with example data which can be download from here.This data was simulated using HAPMAP3 data to mate individuals who already had children to create 301 affected relative pairs (ARPs). The data only contains SNPs from chromosome 6, however in a real analysis we would ideally have data for the whole genome to get better estimates of the prior IBD sharing probabilities and to compare LOD scores across the genome.

We already have PLINK installed on our system so we will not bother ourselves to download it from here, nor will we use the -plink option as it is already set up to run by typing "plink", which is set by default. The KING program (version 2.2.9) on the other hand is not installed, so we shall download it from here, and save the executable file at location /home/me/my-programs/king/ so that KING may be ran by typing /home/me/my-programs/king/king. **Note:** It is very important to download KING version 2.2.9 and not any later versions as we need to use the --homog option which is no longer available in later versions.

We can now run a Regional IBD Analysis (RIA) analyses using the default options for the SNP window size, set to 15 cM, and the SNP window step size, set to 50, by typing:

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o
    resultsRIAExample.dat
```

This will create output similar to the following.

```
RIA: Regional IBD Analysis, v1.1
----------------------------------------------------------------
Copyright 2015-present Richard Howey, GNU General Public License, v3
Research Software Engineering, Newcastle University


Parameters:
Input file: demo-data/exampleRIAData.bed
Output file: resultsRIAExampleDataKing2023.dat
Log file: resultsRIAExampleDataKing2023.log
Using additive and dominance variance model
Start at first SNP with full SNP window
End at last SNP with full SNP window
Window size: 15 cM
Minimum number of SNPs in a window: 100
SNP step size: 50 SNPs
Number of cases: 362
Number of unused controls: 35
Number of SNPs: 29413


Creating list of pruned SNPs using PLINK command:
plink --noweb --indep 50 50 2 --mind 0.01 --maf 0.25 --geno 0.05 --bfile
    exampleRIAData --out tempRIA-priors1-8553 >/dev/null 2>&1


Creating data file to calculate priors using PLINK command:
plink --noweb --bfile exampleRIAData --filter-cases --extract tempRIA-priors1
    -8553.prune.in --make-bed --out tempRIA-priors2-8553 >/dev/null 2>&1


Calculating priors using KING command:
/home/me/my-programs/king/king -b tempRIA-priors2-8553.bed --homog --prefix
    tempRIA-priors2-8553 >/dev/null 2>&1


Number of affected relative pairs (ARPs) in priors: 301


Calculating posteriors (for each SNP window) using KING command:
/home/me/my-programs/king/king -b tempRIA-posterior-8553.bed --homog --prefix
     tempRIA-posterior-8553 >/dev/null 2>&1


15 cM windows - number of SNPs summary statistics:
Mean: 2442.33
Median: 2266
Standard deviation: 606.466
Range: (1458, 3634)


Run time: 10 minutes
```

The commands used by PLINK and KING are output for reference and may be useful if there are any problems. RIA uses several intermediate temporary files beginning with

"tempRIA" and may be lying around if there was a problem and RIA was forced to unexpectedly stop. They should then be carefully deleted if necessary.

The results file `resultsRIAExample.dat` should look as follows:

```
SNP CHR ID CM BP VAR_A VAR_D MLS
741 6 rs11755767 7.52007 2804601 1.583936049 0 2.828838716
791 6 rs7765538 8.14908 3000876 1.544010298 0 2.726726528
841 6 rs744375 8.85487 3294907 1.473726461 0 2.565180586
891 6 rs7763703 9.47382 3433624 1.392624092 0 2.383495002
941 6 rs12526106 10.0032 3690722 1.339150984 0 2.272297645
...
28341 6 rs11755875 188.372 166136328 1.21271465 0.1564407936 2.774211867
28391 6 rs705791 188.972 166288440 1.127065043 0.1950593781 2.693304349
28441 6 rs12154075 189.175 166486261 1.12910064 0.1965589048 2.69542933
28491 6 rs1033725 189.362 166627240 1.125312283 0.1817039321 2.647993295
28541 6 rs6935542 190.191 166817722 0.9200496884 0.3786320582 2.675835187
```

The columns for the results file are as follows:

Column	Description
SNP	The SNP number as it appears in file.
CHR	Chromosome of the SNP.
ID	The name of the SNP.
CM	The centimorgan value of the SNP.
BP	The base pair position of the SNP.
VAR_A	The additive variance parameter.
VAR_D	The dominance variance parameter.
MLS	The maximum-likelihood statistic.

It is not unusual for either VAR_A or VAR_D or even both of these parameters to be equal to 0. In the example data set VAR_D is equal to 0 for most of the SNPs.

In R type:

```
resultsRIAExample<-read.table("resultsRIAExample.dat", header=TRUE)

plot(resultsRIAExample$BP/10^6, resultsRIAExample$MLS, main="Regional IBD
    Analysis", xlab=expression(bp~position~(Mb)), ylab="MLS", ylim=c(0, max(
    resultsRIAExample$MLS)))
```

This will produce the following plot:

As the prior IBD sharing probabilities are always the same, regardless of which subset of SNPs are analysed, it is possible to save a prior file using the `-o-prior` option as follows

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o
    resultsRIAExample.dat -o-prior examplePrior.dat
```
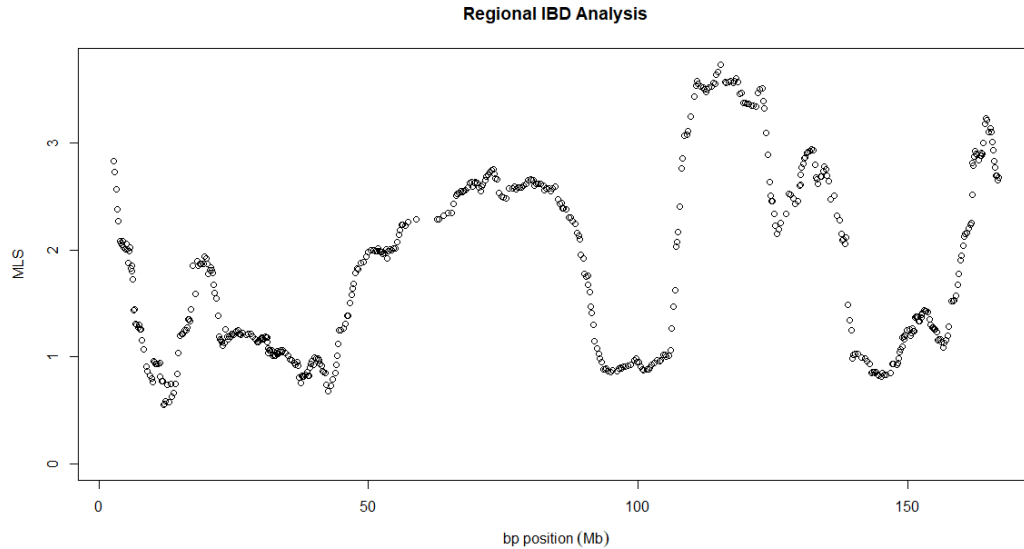
Figure 1: Plot of RIA test results.

or can be done without any further analysis using the `-prior-only` option as follows

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -prior-only
    -o-prior examplePrior.dat
```

The prior can then be used for different analysis using the `-i-prior` option, such as

```
./ria -window-size 2000 -king /home/me/my-programs/king/king -i
    exampleRIAData.bed -o resultsRIAExample-2000.dat -i-prior examplePrior.dat
```

There are no simulated effects in the example data, which would be clearer if results for the whole genome were available. For details on how to interpret results please see Eu-Ahsunthornwattana et al. (2015).

4.1 Parallel processing

Regional IBD Analysis is fairly computationally intensive and so it is natural to want to speed things up a bit by using parallel processing. This can be done by dividing the analysis up using the `-start-snp` and `-end-snp` options, such as

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o
    resultsRIAExample1.dat -i-prior examplePrior.dat -start-snp 3000 -end-snp
    4000
```

Care must be taken when setting the first SNP to ensure there is a full SNP window around the SNP at the center of the SNP window.

If all of the data is to be analysed it is much easy to use the -job option to divide the analysis into a number of analyses. This will automatically set the start and end SNPs. For example, to analyse all of the data in 10 jobs using the previously calculated priors, use

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results1.
    dat -i-prior examplePrior.dat -job 1 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results2.
    dat -i-prior examplePrior.dat -job 2 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results3.
    dat -i-prior examplePrior.dat -job 3 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results4.
    dat -i-prior examplePrior.dat -job 4 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results5.
    dat -i-prior examplePrior.dat -job 5 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results6.
    dat -i-prior examplePrior.dat -job 6 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results7.
    dat -i-prior examplePrior.dat -job 7 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results8.
    dat -i-prior examplePrior.dat -job 8 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results9.
    dat -i-prior examplePrior.dat -job 9 10
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o results10
    .dat -i-prior examplePrior.dat -job 10 10
```

The temporary intermediate files used by RIA depend on the process ID of each job so there is no need to worry about these files interfering with one another. Only the first results file will contain the header making it easy to combine the results into one results file:

```
cat results1.dat results2.dat results3.dat results4.dat results5.dat results6
    .dat results7.dat results8.dat results9.dat results10.dat > allResults.dat
```

The -job option makes it easy to write a simple script with a loop to submit the jobs. Alternatively, if you are using a High Performance Computing (HPC) cluster using the open-source Sun Grid Engine (SGE) scheduler software, then these jobs may be submitted as an array job using something similar to the following script:

```
#!/bin/bash
# execute in current working directory
#$ -cwd
# export local envirnoment
#$ -V
# the number of RIA tasks
#$ -t 1-10
```

```
# execute RIA for each task
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o
    results$SGE_TASK_ID.dat -i-prior examplePrior.dat -job $SGE_TASK_ID 10
```

References

Cordell HJ, Wedig GC, Jacobs KB, Elston RC. 2000. Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 66:1273–1286.

Eu-Ahsunthornwattana J, Howey R, Cordell HJ. 2015. Regional IBD Analysis. TBA .

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, M CW. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.