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.

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 Manichaikul et al. (2010).

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. These probabilities give the *posteriors*.

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).

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 *.cpp -o ria
```

3. It is also necessary to having working copies of the programs PLINK and KING installed.

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 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 mydata.bed
```

Typing `ria` with no options will output usage details:

```
RIA: Regional IBD Analysis, v1.00
------------------------------------------------------------
Copyright 2015 Richard Howey, GNU General Public License, v3
Institute of Genetic Medicine, Newcastle University

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

Options:
  -window-size n      -- set window SNP size to n SNPs
  -ws-total-snps t    -- set window SNP size given total SNPs chr1 to chr22
  -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
  -job a m            -- job number a of m
  -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 1000
  -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 2000
-step-size 50
-i mydata.bed
-o myResults.bed
```

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 2000

# 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
```

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 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`.

We can now run a Regional IBD Analysis (RIA) analyses using the default options for the SNP window size, automatically set depending on the given data, 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, v1.00
-------------------------------------------------------------
```

```
Copyright 2015 Richard Howey, GNU General Public License, v3
Institute of Genetic Medicine, Newcastle University

Parameters:
Input file: exampleRIAData.bed
Output file: resultsRIAExample.dat
Log file: resultsRIAExample.log
Using additive and dominance variance model
Start at first SNP with full SNP window
End at last SNP with full SNP window
SNP step size: 50 SNPs
Number of cases: 362
Number of unused controls: 35
Number of SNPs: 29413
SNP window size: 2010 SNPs, given by estimated total SNPs in genome (chrs 1
   to 22): 446663.275146

Creating list of pruned SNPs using PLINK command:
plink --noweb --indep 50 50 2 --mind 0.01 --maf 0.25 --nonfounders --geno
   0.05 --bfile exampleRIAData --out tempRIA-priors1-2499 >/dev/null 2>&1

Creating data file to calculate priors using PLINK command:
plink --noweb --bfile exampleRIAData --nonfounders --filter-cases --extract
   tempRIA-priors1-2499.prune.in --make-bed --out tempRIA-priors2-2499 >/dev/
   null 2>&1

Calculating priors using KING command:
/home/nrajh/code-other/king -b tempRIA-priors2-2499.bed --homo --prefix
   tempRIA-priors2-2499 >/dev/null 2>&1

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

Calculating posteriors (for each SNP window) using KING command:
/home/nrajh/code-other/king -b tempRIA-posterior-2499.bed --homo --prefix
   tempRIA-posterior-2499 >/dev/null 2>&1


Run time: 4 minutes and 59 seconds
```

The SNP window size is set automatically if not set by the `-window-size` option. If data for chromosomes 1 to 22 are given then the SNP window size is set to the total number of SNPs in chromosomes 1 to 22 multiplied by 450/100,000. If data for only some chromosomes are given then a rough estimate is made of the total number of SNPs in chromosomes 1 to 22, and this estimate is then used to set the SNP window size. In example data consists of data for chromosome 6 only, and so an estimate is made of the total SNPs to set the window size. Suppose that we know that there are in fact 500000 SNPs in chromosomes 1 to 22 then this number can be set to estimate the SNP window size using the -ws-total-snps option:

```
./ria -king /home/me/my-programs/king/king -i exampleRIAData.bed -o
   resultsRIAExample500000.dat -ws-total-snps 500000
```

This results in a SNP window size of 2250 and gives very similar results. If the data is split across different files then it would be best to use the same SNP window size either by using the -ws-total-snps option or the -window-size option.

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. These should be carefully deleted if necessary.

The results file resultsRIAExample.dat should look as follows:

```
SNP CHR ID BP VAR_A VAR_D MLS
1006 6 rs10458166 3957928 1.338110792 0 2.141627931
1056 6 rs1335277 4161931 1.300657856 0 2.057961326
1106 6 rs2225369 4392744 1.342525466 0 2.118241
1156 6 rs433874 4599981 1.418617946 0 2.237598337
1206 6 rs233477 4741997 1.339948243 0 2.099292049
...
28206 6 rs717602 165727097 1.252687418 0.03895511046 2.520487693
28256 6 rs635547 165852256 1.186978244 0.1065481174 2.517938126
28306 6 rs3008049 165983607 1.109034121 0.165932504 2.465117525
28356 6 rs9348050 166183478 1.044282802 0.1992880298 2.387806547
28406 6 rs697491 166361018 0.9595321341 0.2864577139 2.405823963
```

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.
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")
```

This will produce the following plot:

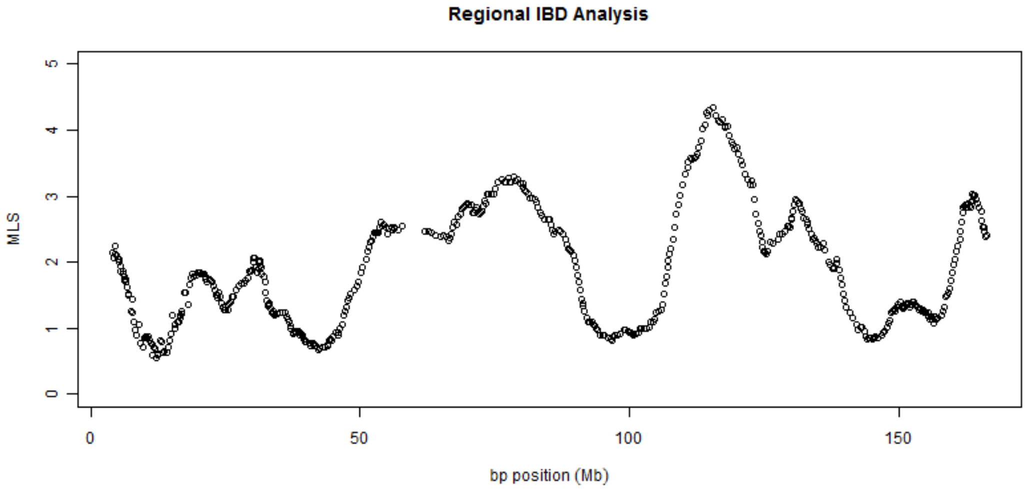

Figure 1: Plot of RIA test results.

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
```

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
```

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.