Computer Practical Exercise on Family-based Association using FaST-LMM, PLINK and R

Overview

Purpose

In this exercise you will be carrying out association analysis of data from a mini genome-wide association study. The data comes from families (related individuals) measured for a quantitative trait of interest. The purpose is detect which (if any) of the loci are associated with the quantitative trait.

Methodology

We will use the linear mixed model approach implemented in FaST-LMM and (for comparison) standard linear regression in PLINK.

Program documentation

PLINK documentation:

PLINK has an extensive set of docmentation including a pdf manual, a web-based tutorial and web-based documentation:

Original PLINK (1.07): http://zzz.bwh.harvard.edu/plink/

New PLINK (1.90): https://www.cog-genomics.org/plink2

R documentation:

The R website is at http://www.r-project.org/

From within R, one can obtain help on any command `xxxx` by typing `help(xxxx)'

FaST-LMM documentation:

Documentation can be downloaded together with the FaST-LMM program from

http://research.microsoft.com/en-us/downloads/aa90ccfb-b2a8-4872-ba00-32419913ca14/

Data overview

We will be using family data consisting of 498 individuals typed at 134,946 SNPs. All individuals have measurements of a quantitative trait of interest. You can assume that appropriate quality control (QC) checks on SNPs and individuals have been carried out prior to the current analysis i.e. the data set is already QC-ed.

Appropriate data

Appropriate data for this exercise is genome-wide genotype data for related and/or apparently unrelated individuals. Genome-wide data is required in order to estimate relationships between people and allow for relatedness in the analysis. The individuals should be phenotyped for either a dichotomous trait or a quantitative trait of interest.

Instructions

Data files

The data is in PLINK binary-file format. Check you have the required files by typing:

```
ls -l
```

You should find 3 PLINK binary-format files in your directory: `quantfamdata.bed`, `quantfamdata.bim` and `quantfamdata.fam`. The file `quantfamdata.bed` is the binary genotype file which will not be human readable. The file `quantfamdata.bim` is a map file. You can take a look at this (e.g. by typing `more quantfamdata.bim`). The file `quantfamdata.fam` gives the pedigree structure in a format that is compatible with the binary genotype file. You can take a look at this (e.g. by typing `more quantfamdata.fam`). Note this file is the same as the first six columns of a standard pedigree file, with the last column giving each individual's quantitative trait value.

Step-by-step instructions

1. Analysis in PLINK

To start with, we will use PLINK to perform a test equivalent to linear regression analysis, without worrying about the relatedness between individuals:

```
plink --bfile quantfamdata --assoc --out plinkresults
```

A copy of the screen output is saved in the file `plinkresults.log`. The association results are output to a file `plinkresults.qassoc`. Take a look at this file. Each line corresponds to the results for a particular SNP. Each line contains the following columns:

```
     CHR      Chromosome number
     SNP      SNP identifier
     BP       Physical position (base-pair)
     NMISS    Number of non-missing genotypes
     BETA     Regression coefficient
     SE       Standard error
     R2       Regression r-squared
     T        Wald test (based on t-distribtion)
     P        Wald test asymptotic p-value
```

The most useful columns are T (the test statistic) and its p value (P).

To visualise these results properly we will use R. Open up a new terminal window, move to the directory where you performed this analysis, and start R (by typing `R`).

Now (within R) read in the data by typing:

```
res1<-read.table("plinkresults.qassoc", header=T)
```

This reads the results into a dataframe named "res1". To see the top few lines of this dataframe, type:

```
head(res1)
```

The data frame has 134,946 lines, one for each SNP. It would be very laborious to go through and look at each line by eye. Instead we will plot the results for all chromosomes, colouring each chromosome differently. To do this we need to first read in from an external file some special functions for creating such ``Manhattan" plots:

```
source("qqmanHJCupdated.R")
```

Then we use the folllowing command to actually make the plot:

```
manhattan(res1, pch=20, suggestiveline=F, genomewideline=F, ymin=2, cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
```

Be warned, this may take some time to plot.

Visually it looks like there may be significant results on chromosomes 6 and 12, and possibly on chromosome 5 as well. One way to assess the significance of the results, in light of the large number of tests performed, is to use a Q-Q plot. To plot a Q-Q plot for these P values, type:

```
qq(res1$P)
```

What one would hope to see is most of the values lying along the straight line with gradient 1, indicating that most results are consistent with the null hypothesis of no association. However, one would also hope to see a few high values at the top that depart from the straight line, which are hopefully true associations.

Our results seem fairly consistent with this expectation, but there may be a little bit of inflation (i.e. a slope slightly bigger than 1) due to relatedness between individuals. To calculate the genomic control inflation factor, we first convert the P values to chi-squared test statistics on 1df, and then use the formula from Devlin and Roeder (1999):

```
chi<-(qchisq(1-res1$P,1))
lambda=median(chi)/0.456
lambda
```

You should find a slightly inflated value (lambda=1.10)

2. Analysis in FaST-LMM

Now we will try re-running the analysis using FaST-LMM, which estimates and accounts for the relatedness between individuals. Go back to the window where you ran PLINK and run FaST-LMM as follows:

```
fastlmmc -bfile quantfamdata -pheno quantfamdata.fam -mpheno 4 -bfileSim quantfamdata -ML -out FLMMresults
```

Here we use the `-bfile quantfamdata` command to tell the program the name (stem) of the files with the input genotype data containing the SNPs to be tested for association, and the `-bfileSim quantfamdata` command to tell the program the name of the files containing the SNPs to be used for estimating relatedness. Here we just use the same files both times, but FaST-LMM would allow us to use different files for these two operations if we prefer.

The command `-pheno quantfamdata.fam -mpheno 4` tells FaST-LMM to read the phenotype data in from the file `quantfamdata.fam` , using the 4th phenotype column (not including the two first columns which give the family and person IDs). The `-ML` command tells FaST-LMM to use maximium likelihood estimation which we recommend as opposed to restricted maximium likelihood (REML). The command `-out FLMMresults` tells FaST-LMM the name to use for the output file.

Take a look at the results file. FaST-LMM automatically orders the results by significance.

Now go back to your R window and read the results into R:

```
res2<-read.table("FLMMresults", header=T)
```

Check the column names by typing:

```
head(res2)
```

The P value is in a column called ``Pvalue". Remember FaST-LMM has automatically ordered the results by significance, so these top few rows will show the most significant results.

First let us check the genomic control inflation factor. We convert the P values to chi-squared test statistics on 1df, and then use the formula from Devlin and Roeder (1999):

```
chi<-(qchisq(1-res2$Pvalue,1))
lambda=median(chi)/0.456
lambda
```

You should find a less inflated value (lambda=0.99) than we found previously with PLINK.

To plot Manhattan and Q-Q plots you can use similar commands to before, but the columns need to be named appropriately. The easiest thing is to make a new smaller dataframe containing the required data:

```
new<-data.frame(res2$SNP, res2$Chromosome, res2$Position, res2$Pvalue)
names(new)<-c("SNP", "CHR", "BP", "P")
head(new)
```

Now you can plot the Q-Q plot:

```
qq(new$P)
```

And the Manhattan plot:

```
manhattan(new, pch=20, suggestiveline=F, genomewideline=F, ymin=2, cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
```

The significant effects on chromosomes 6 and 12 are still easily visible. In fact, this is simulated data, and these signals do correspond correctly to the positions of the underlying causal variants.

Answers

How to interpret the output

Interpretation of the output is described in the step-by-step instructions. In general, the output will consist of a likelihood-ratio or chi-squared test for whatever you are test you are performing, and regression coefficients or odds ratio estimates for the predictor variables in the current model. Please ask if you need help in understanding the output for any specific test.

Comments

Advantages/disadvantages

PLINK is useful for data management and analysis of genome-wide association data. FaST-LMM is more appropriate for analysis of related individuals, or for correcting for population stratification in apparently unrelated individuals.

Other packages

Other packages that can implement a similar analysis to FaST-LMM include EMMAX, GEMMA, MMM, GenABEL, Mendel.

References

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies Nat Methods 8(10):833-835.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81:559-575.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: PLINK, R, FaST-LMM
Last updated: 09/01/2023 11:19:05