Interaction analysis using PLINK and CASSI

Overview

Purpose

In this exercise you will be performing association analysis and testing for interaction effects using case/control data.

Methodology

The methodology used includes logistic regression in PLINK and CASSI, as well as some related alternative approaches.

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) (which has arguably clearer documentation):
http://zzz.bwh.harvard.edu/plink/

New PLINK (1.90) (which includes documentation on new additional features):
https://www.cog-genomics.org/plink2

CASSI documentation:

CASSI documentation is available from:

http://www.staff.ncl.ac.uk/richard.howey/cassi/downloads.html

Exercise

Data overview

The data consists of simulated genotype data at 100 SNP loci, typed in 2000 cases and 2000 controls. The data has been simulated in such a way that the first

five SNPs have some relationship with disease, whereas the remaining 95 SNPs have no effect on disease outcome.

The complication with these data is that SNPs 1 and 2 have been simulated in such a way that they show no marginal association with the disease: their association will only be visible when you look at both SNPs in combination. SNPs 3-5 have been simulated to only have an effect on disease when an individual is homozygous at all three of these loci. Although potentially this could lead to marginal effects at the loci, formally this corresponds to a model of pure interaction, with no main effects, at these 3 SNPs.

Appropriate data

Appropriate data for this exercise is genotype data for a set of linked or unlinked loci typed in a group of unrelated affected individuals (cases) and in a group of unaffected or randomly chosen individuals from the same population (controls).

All the programs will deal with much larger numbers of loci than the 100 SNPs considered here. PLINK, in particular, was specifically designed for the analysis of large numbers of loci e.g. generated as part of a genome-wide association study.

Instructions

Data format

The data for the 100 SNPs simcasecon.ped is in standard linkage pedigree file format, with columns corresponding to family id, subject id (within family), father's id, mother's id, sex (1=m, 2=f), affection status (1=unaffected, 2=affected) and one column for each allele for each locus genotype. Note that since this is case/control rather than family data, there is only one individual per family and everyone's parents are coded as unknown.

PLINK requires an additional map file simcasecon.map describing the markers (in order) in the pedigree file. The PLINK-format map file contains exactly 4 columns:

```
chromosome (1-22, X, Y or 0 if unplaced)
rs number or snp identifier
Genetic distance (morgans) (not often used - so can be set to 0)
Base-pair position (bp units)
```

Take a look at the data files, and check that you understand how the data is coded. Then (if necessary) save the files as .txt files to the appropriate directory (folder) on your computer.

Step-by-step instructions

1. Analysis in PLINK

Move to the directory where you have saved the data files.

To carry out a basic association analysis in PLINK, type

```
plink --ped simcasecon.ped --map simcasecon.map --assoc
```

Here the `--ped xxxx` command tells PLINK that the name of the pedigree file is `xxxx` and the `--map yyyy` command tells PLINK that the name of the map file is `yyyy`. The `--assoc` command tells PLINK to perform a basic allele-based chisquared association test.

PLINK outputs some useful messages (you should always read these to make sure they match up with what you expect!) and outputs the results to a file `plink.assoc` .

Take a look at the file `plink.assoc` (e.g. by typing `more plink.assoc`). For each SNP the following columns of results are reported:

CHR	**Chromosome**
SNP	**SNP ID**
BP	**Physical position (base-pair)**
A1	**Minor allele name (based on whole sample)**
F_A	**Frequency of this allele in cases**
F_U	**Frequency of this allele in controls**
A2	**Major allele name**
CHISQ	**Basic allelic test chi-square (1df)**
P	**Asymptotic p-value for this test**
OR	**Estimated odds ratio (for A1, i.e. A2 is reference)**

Does there appear to be evidence of association at any of the five "true" loci? What about the 95 null loci?

Try performing a genotype-based (rather than an allele-based) analysis in PLINK and take a look at the results by typing the following 3 commands:

```
plink --ped simcasecon.ped --map simcasecon.map --model
head -1 plink.model
grep GENO plink.model
```

Again, does there appear to be evidence of association at any of the five "true" loci? What about the 95 null loci?

To test for pairwise epistasis in PLINK, the fastest option is to use the `--fast-epistasis` command:

```
plink --ped simcasecon.ped --map simcasecon.map --fast-epistasis
```

Formally, this tests whether the OR for association between two SNPs differs between cases and controls, which can be shown to appriximate a logistic regression based test of interaction between the SNPs. Results can be found in the file `plink.epi.cc`. Only pairwise interaction tests with p <= 0.0001 are reported (otherwise, for genome-wide studies, there would be too many results to report, given the large number of pairwise tests performed).

Take a look at the file `plink.epi.cc`. You should find a very significant interaction between SNPs 1 and 2, and a less significant iteraction between SNPs 15 and 77. Since this is simulated data, we know that this less significant result is a false positive.

A more powerful test for SNPs that are not in LD with one another (i.e. that are

not too close to one another, in terms of their genomic location) is to additionally use the `--case-only` option:

```
plink --ped simcasecon.ped --map simcasecon.map --fast-epistasis --case-only
```

Results can be found in the file `plink.epi.co` . Again only pairwise interaction tests with p <= 0.0001 are reported. You should again find a very significant interaction between SNPs 1 and 2 (even more significant than previously, owing to the increased power with a case-only test).

A problem with the --fast-epistasis test is that it can be affected by LD between the SNPs (although only the case-only test is seriously affected). A more accurate test is to carry out logistic regresion by using the slower `--epistasis` command:

```
plink --ped simcasecon.ped --map simcasecon.map --epistasis
```

Results can again be found in the file `plink.epi.cc` (which will now have been overwritten). You can see that again the interaction between SNPs 1 and 2 remains highly significant (p=1.22E-63), together with just one other (false positive) interaction between SNPs 15 and 77.

Since the `--epistasis` option is slower, but most accurate, for genome-wide studies it might be sensible to first to screen for interactions using the `--fast-epistasis` command, but then confirm any findings using the `--epistasis` command on the smaller set of detected SNPs.

---

2. Analysis in CASSI

We will also compare our PLINK results with those obtained using the CASSI program, which implements a variety of tests including linear and logistic regression, and an improved Joint Effects (JE) test of pairwise interaction as described in Ueki and Cordell (2012). First we need to convert our data to PLINK binary format:

```
plink --ped simcasecon.ped --map simcasecon.map --make-bed --out simbinary
```

This should create PLINK binary format files `simbinary.bed`, `simbinary.bim` and `simbinary.fam`. Then we use the CASSI program with the input file `simbinary.bed` to perform pairwise interaction tests at all pairs of loci. (By default, only those pairs of SNPs showing interaction with a p-value < 0.0001 are output, though this can be changed if desired).

We start by using logistic regression. The logistic regression test in CASSI is essentially the same as the `--epistasis` test in PLINK, except that CASSI uses a likelihood ratio test rather than the asymptotically equivalent Wald (?) test used by PLINK. CASSI also has the advantage of allowing covariates into the analysis, if desired.

```
cassi -lr -i simbinary.bed
```

Take a look at the output file `cassi.out` The most important columns are the first 4 columns (listing the SNP numbers/names) and the last 4 columns listing the log odds ratio, its standard error, the likelihood ratio chi-squared test statistic and its p-value. It can be quite hard to work out which column is which, so we suggest you start up R by typing

```
R
```

and then read in and look at the results by typing

```
results<-read.table("cassi.out", header=T)
results
```

You can see that SNPs 1 and 2 show a very strong pairwise interaction (p=5.94E-72), which is actually a bit more significant than the result from PLINK (p=1.22E-63). We also still detect the false positive interaction between SNPs 15 and 77.

Now try using the Joint Effects (JE) test, telling CASSI to use the output filename `cassiJE.out`

```
cassi -je -o cassiJE.out -i simbinary.bed
```

Take a look at the output file `cassiJE.out`. The most important columns are the first 4 columns (listing the SNP numbers/names) and the last 4 columns listing the case/control and case-only interaction test chi-squareds and p-values. Again it can be quite hard to work out which column is which, so we suggest you read in and look at the results in R:

```
resultsJE<-read.table("cassiJE.out", header=T)
resultsJE
```

You can see that SNPs 1 and 2 show a very strong pairwise interaction (Case-Con test p-value JE_CC_P=1.67e-129; Case-Only test p-value JE_CO_P=1.71e-274). Interestingly we also detect, albeit at lower significance levels, the (true) pairwise interactions between SNPs 3 and 4 and between SNPs 4 and 5. We also detect two false positive interactions, between SNPs 15 and 77, and between SNPs 31 and 100.

Answers

Interpretation of output

Answers and interpretation of the output are described in the step-by-step instructions. Please ask if you need help in understanding the output for any specific test.

Comments

Advantages/disadvantages

PLINK and CASSI are designed for genome-wide studies, allowing the inclusion of many thousands of markers. Analysis in a standard statistical package does not generally allow so many markers, but may have some advantage of allowing a lot of extra flexibility with regards to the models and analyses performed e.g. it easy

to include additional predictor variables such as environmental factors, gene-environment interactions etc. However, you are required to know or learn how to use the package in order to gain that extra flexibility, and to produce reliable results.

Study design issues

With case/control data it is relatively easy to obtain large enough sample sizes to detect small genetic effects. However, detection of interactions generally requires much larger sample sizes.

Other packages

Logistic regression analysis for detection of interactions can be performed in most statistical packages such as R, Stata, SAS, SPSS. Alternative Bayesian Epistasis mapping approaches are available in the BEAM (Zhang et al. 2007; Zhang 2011) or BIA software packages.

Several packages are available for implementing different data-mining and machine-learning approaches for detecting interactions or detecting association allowing for interaction. See Cordell (2009) and other references below for more details.

References

Cordell HJ (2009) Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10(6):392-404.

Y Chung and S Y Lee and R C Elston and T Park (2007) Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics 23:71-76.

L W Hahn and M D Ritchie and J H Moore (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions Bioinformatics 19:376-382.

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.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF and Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138-147.

Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genetics 8(4):e1002625.

Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167-1173.

Zhang Y (2011) A novel Bayesian graphical model for genome-wide multi-SNP association mapping. Genet Epidemiol 36: 36-47.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: PLINK, CASSI
Last updated: 01/17/2020 12:35:48