

CASSI: Genome-Wide Interaction Analysis Software

Contents

1	Introduction	3
2	Installation	3
3	Using CASSI	3
3.1	Options	4
3.2	Parameter file	5
4	Joint Effects	6
4.1	Options	6
4.2	Input Files	6
4.3	Output Files	7

1 Introduction

CASSI (Contrived Acronym of Software for SNP Interactions) is a C++ program written to analyse SNP-SNP interactions, analysing a choice of SNPs from two given SNP windows (possibly from different pedigree files). Each pair of SNPs whose interaction test passes a given significance level is returned in the output file with extra information for the performed test. The program only accepts PLINK binary files (`.bed`) in order to perform the calculations as efficiently as possible.

The only test that is currently available is the *joint effects* test Ueki and Cordell (2012), details of this test are given in section 4. Further interaction tests are planned for CASSI so watch this space!

CASSI is copyright, 2012 Richard Howey, GNU General Public License, version 3.

2 Installation

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

1. Download the code from the download page.
2. Compile it by typing something like the following:

```
g++ -O3 *.cpp -o cassi
```

3. Start using CASSI!

3 Using CASSI

Basic usage of CASSI is to provide it with one binary PLINK format pedigree file:

```
./cassi myfile.bed
```

This requires that the corresponding `.bim` and `.fam`, files are also available. A text PLINK pedigree file, `.ped`, with corresponding map file, `.map`, may be used to create a binary file using PLINK as follows:

```
plink --noweb --file mydata --make-bed --out myfile
```

This will create the binary pedigree file, `myfile.bed`, map file, `myfile.bim`, and family file, `myfile.fam` required for use with CASSI.

Executing CASSI as above will perform SNP interaction tests for every pair of distinct SNPs in the `.bed` file using the default options. The results file will record every pair of SNPs that satisfy a given significance level with extra information for the performed test. A log file is also created recording the same information that is output to the screen, showing the used options and summary statistics of the data. It is unlikely that you will want to use the default options and typing `./cassi` with no options will output the available options.

CASSI is executed as follows:

```
./cassi [options] file.bed
```

or

```
./cassi parameterfile.pf [file.bed]
```

3.1 Options

The basic options for CASSI are as follows:

Option	Description
-snp1 a1 a2	First SNP window, a1 = Start SNP number, a2 = End SNP number
-snp2 b1 b2	Second SNP window, b1 = Start SNP number, b2 = End SNP number
-i file.bed	Input file
-i2 file2.bed	Second (optional) input file for second SNP window
-o file.out	Results output file
-log file.log	Log file
-max m	Maximum number of results, to safeguard accidentally outputting half a trillion results.
-so	suppress output to screen

The SNP numbers “a1” and “a2” etc. refer to the position the SNP appears in the map file (.bim).

For example, to use these options to analyse SNPs from SNP number 1 to SNP number 60 against SNPs from SNP number 50 to SNP number 100 using binary pedigree file mydata.bed type the following:

```
./cassi -snp1 1 60 -snp2 50 100 mydata.bed
```

This will output details of the analysis and will look something like the following:

```
CASSI: SNP interaction analysis software, v1.0
```

```
-----
```

```
Copyright 2012 Richard Howey, GNU General Public License, v3  
Institute of Genetic Medicine, Newcastle University
```

```
Parameters:
```

```
Input file: mydata.bed
```

```
Output file: cassi.out
```

```
Start SNP of first SNP window: 1
```

```
End SNP of first SNP window: 60
```

```
Start SNP of second SNP window: 50
```

```
End SNP of second SNP window: 100
```

```
Maximum no. of results: 1000000
```

```
Test Statistic: Joint Effects
```

```
P-value threshold for case/control results: 0.0001
```

P-value threshold for case only results: 0.0001

Data Summary Statistics:

Number of SNPs: 100

Number of subjects: 4686

Number of cases: 1748 (37.3026%)

Number of controls: 2938 (62.6974%)

Number of results found: 80

Run time: less than one second

To do the above analysis where the second SNP file is given in a different pedigree file type the following:

```
./cassi -snp1 1 60 -snp2 50 100 -i mydata.bed -i2 mydata2.bed
```

Options that are specific to the joint effects test are:

Option	Description
-thcc t	P-value threshold for case/control test (set to 0 for no output)
-thco t	P-value threshold for case only test (set to 0 for no output)
-th t	P-value threshold for either test (set to 0 for no output)

These options set the p-value thresholds for the joints effects SNP interaction test to give a result that is significant enough to record in the results file. For more details see section 4

The default options for CASSI are:

Option	Description
-snp1 a1 a2	All SNPs in pedigree file
-snp2 b1 b2	All SNPs in pedigree file (or 2nd pedigree file if given)
-o	cassi.out
-log	cassi.log
-th	0.0001
-max m	1000000 (10^6)

3.2 Parameter file

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

```
./cassi 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:

```
-snp1 1 60
-snp2 50 100
-i mydata.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 to perform the next analysis given above:

```
This is the first SNP window
-snp1 1 60
```

```
This is the second SNP window
-snp2 50 100
```

```
This is the pedigree file for the first SNP window
-i mydata.bed
```

```
This is the pedigree file for the second SNP window
-i2 mydata2.bed
```

```
I might try this threshold later
#-th 0.00001
```

4 Joint Effects

For full details of the joints effects test and the accompanying notation please refer to the manuscript by Ueki and Cordell (2012).

4.1 Options

The options that are specific to the joint effects test are:

Option	Description
-thcc t	P-value threshold for case/control test (set to 0 for no output)
-thco t	P-value threshold for case only test (set to 0 for no output)
-th t	P-value threshold for either test (set to 0 for no output)

CASSI performs both the case-only test and case-control test for each pair of tested SNPs. The results of the test are recorded for the pair of SNPs if *either* p-value threshold is met. Both thresholds may be set simultaneously by using the `-th` option. The default value for both thresholds is 0.0001.

For other basic options see section 3.1.

4.2 Input Files

The input files should be given in the binary PLINK format as explained in section 3.

4.3 Output Files

The output file from running a joint effects SNP interaction analysis using CASSI will look something like the following:

```
SNP1 SNP1_Name SNP2 SNP2_Name Case_logOR Case_SE Con_logOR Con_logSE
Case-Con_Chisq Case-Con_Pval Case-Only_Chisq Case-Only_Pval
4 rs123456 56 rs2345 4.13044 0.946829 3.96474 0.983261 0.00910349 0.923987 15.2945 9.
4 rs123456 73 rs2365 4.19374 1.09557 5.31239 1.01136 0.0720141 0.788426 21.1098 4.337
4 rs123456 74 rs342345 3.93546 1.02951 5.40301 0.878292 5.89417e-06 0.998063 16.4154
...
```

The first line is a header labelling the columns of the results file. The SNP number and SNP names for the pair of SNPs are given firstly followed by values calculated from the joint effects tests. Results are given for both the case-only test and for the case-control test regardless of which p-value was significant enough to record the pair of SNPs.

Case_logOR is the log odds for the cases and is given by $\tilde{\lambda}_A$. If the alternative test is used this is given by $\tilde{\mu}_A$.

Case_SE is the standard error for $\tilde{\lambda}_A$ and is given by $\tilde{\nu}_A$. If the alternative test is used the standard error for $\tilde{\mu}_A$ is recorded, given by $\tilde{\nu}_{\mu_A}$.

Con_logOR is the log odds for the controls and is given by $\tilde{\lambda}_N$. If the alternative test is used this is given by $\tilde{\mu}_N$.

Con_SE is the standard error of the controls test and is given by $\tilde{\nu}_N$. If the alternative test is used the standard error for $\tilde{\mu}_N$ is recorded, given by $\tilde{\nu}_{\mu_N}$.

Case-Con_Chisq is the χ^2 test statistic with one degree of freedom for the case-control test.

Case-Con_Pval is the corresponding p-value for the case-control test.

Case-Only_Chisq is the χ^2 test statistic with one degree of freedom for the case-only test.

Case-Only_Pval is the corresponding p-value for the case-only test.

References

Ueki M, Cordell HJ. 2012. Improved statistics for genome-wide interaction analysis. PLoS Genet 8:e1002625.