SnipSnip: GWAS for Poorly Tagged Data using Multiple SNPs

Contents

1 Introduction

The SnipSnip program is the implementation of a GWAS method to detect causal variants using poorly tagged data by using multiple SNPs in low LD with the causal variant. Genome-wide association studies (GWAS) allow the detection of non-genotyped disease causing variants through the testing of nearby genotyped SNPs that are in strong linkage disequilibrium (LD) with the causal variant. This approach is naturally flawed when there are no genotyped SNPs in strong LD with the causal variant. There may, however, be several genotyped SNPs in weak LD with the causal variant that, when considered together, provide equivalent information. This observation provides the motivation for the popular (but computationally intensive) imputation-based approaches that are often used.

SnipSnip is designed for the scenario where there are several genotyped SNPs in weak LD with a causal variant. Our approach proceeds by selecting, for each genotyped "anchor" SNP, a nearby genotyped "partner" SNP (chosen, on the basis of a specific algorithm we have developed, to be the optimal partner). These two SNPs are then used as predictors in a linear or logistic regression analysis, in order to generate a final significance test associated with the anchor SNP.

SnipSnip is designed for use with **unrelated individuals** either in a case-control analysis or a quantitative trait analysis.

Our method, in some cases, potentially eliminates the need for more complex methods such as sequencing and imputation or haplotype analysis, and provides a useful additional test that may be used with existing GWAS data to identify genetic regions of interest.

For an example application of SnipSnip see Chen et al. (2015).

1.1 Program information and citation

For details concerning the methodology of the SnipSnip GWAS, please see the accompanying manuscript Howey and Cordell (2014).

The program SnipSnip is written in C++ and executables are available for Linux and Windows from the download page, as well as the source code.

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 download page.
2. Compile it by typing something like the following:

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

3. Start analysing your data with SnipSnip!

3 Using SnipSnip

3.1 Basic Usage

The program SnipSnip takes a PLINK binary pedigree file as input. Basic usage of the program is given by typing:

```
./snipsnip -o myresults.dat mydata.bed
```

3.2 Options

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

```
SnipSnip: Imputation without imputation, v1.1
---------------------------------------------
Copyright 2013 Richard Howey, GNU General Public License, v3
Institute of Genetic Medicine, Newcastle University

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

Options:
  -window-size n    -- fix window at n SNPS, n must be even
  -window-size-bp x -- size of window, x, in kB
  -start a          -- start analysis from SNP number a
  -start-end a b    -- start and end analysis from SNP numbers a to b
  -i file.bed       -- input binary pedigree file, file.bed
  -o results.dat    -- output results file, results.dat
  -log results.log  -- log filename, results.log
  -covar covars.dat -- covariate filename, covars.dat
  -covar-number no  -- covariate number, no
  -covar-name na    -- covariate name, na
  -linear           -- use linear regression
  -mqtv x           -- missing quantitive trait value for linear regression
  -dominant         -- use dominant correlation partner metric
  -recessive        -- use recessive correlation partner metric
  -lr               -- perform standard logistic(linear) regression tests
  -excsnp bp        -- exclude SNP(base pair bp) as partner
  -so               -- suppress output to screen

Default Options in Effect:
  -window-size 10
```

```
-o snipsnipResults.dat
```

3.3 Parameter file

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

```
./snipsnip myparameters.pf
```

The parameter file should be a text file with one option written on each line. For example, to perform an analysis with a SNP window of size 12, perform test for SNPs 100 to 200, include single SNP logistic regression results, with input file `mydata.bed` and output file `myresults.dat` the file `myparameters.pf` would be as follows:

```
-window-size 12
-start-end 100 200
-lr
-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:

```
I will use this window size
-window-size 12

Must remember to analysis other SNPs later
-start-end 100 200

Check single SNP logistic regression results also
-lr

This is my data
-i mydata.bed

Output the results here
-o myresults.dat

When I run lots of things I will suppress the output to screen
#-so
```

3.4 Input

SnipSnip takes standard PLINK binary pedigree files, `.bed`, as input. 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 SnipSnip.

3.5 Output

The main results file is given by a text file where each row gives the results for each SNP. For example, using the default options gives the follows:

```
SNP CHR ID BP PARTNER_ID PARTNER_BP CORRELATION SCORE FIT_STATUS CHISQ P
1 0 rs7112558 5569598 rs11038270 5572829 0.1875740 84.85855 Y 1.56046028 0.2115978
2 0 rs7123372 5569768 rs12786429 5570176 0.5597501 68.06836 Y 0.39244220 0.5310185
...
```

The columns of the results file, which will differ depending on the chosen options, are as follows:

Column	Description
SNP	The SNP number (of the anchor SNP) as it appears in file.
CHR	Chromosome of the anchor SNP.
ID	The name of the anchor SNP.
BP	The base pair position of the anchor SNP.
PARTNER_ID	The name of the partner SNP.
PARTNER_BP	The base pair position of the partner SNP.
CORRELATION	The correlation (r^2) between the anchor SNP and partner SNP.
SCORE	The score (0-100) between the anchor SNP and partner SNP. High scores are be
FIT_STATUS	A "Y" indicates that, yes, the model fitted with no problems. An "N" indicates
CHISQ	The χ^2 test statistic with one degree of freedom from performing a likelihood rat
FSTAT	The F test statistic with 1 and *number of subjects*-3 degrees of freedom from per
P	The p-value for the test of association of the anchor SNP.
FIT_STATUS_LR	The fit status for single SNP logistic (or linear) regression at the anchor SNP.
CHISQ_LR	The χ^2 test statistic with one degree of freedom for single SNP logistic regressio
FSTAT_LR	The F test statistic with 1 and *number of subjects*-3 degrees of freedom for single
P_LR	The p-value for the test of association of the anchor SNP using single SNP logist

4 SnipSnip GWAS

A genome-wide association study (GWAS) can be performed using SnipSnip where each SNP is considered in turn as the *anchor* SNP and is tested for association with help from the *partner* SNP, see figure 1.

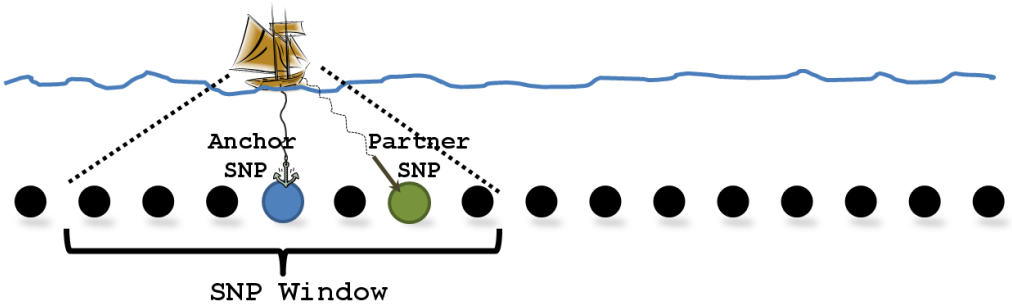

Figure 1: Diagram illustrating the anchor SNP and partner SNP with a SNP window of size 6.

4.1 Anchor SNP

The anchor SNP is simply the SNP that is being tested for association.

4.2 Partner SNP

The partner SNP is chosen from a SNP window surrounding the anchor SNP and is the SNP with the "best" (not highest) LD with the anchor SNP. See section 5 for more infomation on how the partner SNP is chosen.

4.3 AI Test

The artificial-imputation (AI) test is a likelihood ratio test comparing a logistic regression model with the partner SNP only against one with both the anchor SNP and partner SNP. The test produces a χ^2 test statistic with one degree of freedom and tests the significance of the anchor SNP whilst conditioning for the partner SNP. An equivilent linear regression test is also possible.

This is essentially a simple test and is only effective due to the manner in which the partner SNP is chosen.

For more details concerning the methodology, please read the accompanying manuscript Howey and Cordell (2014).

5 Partner SNP Metrics

The partner SNP is chosen using the correlation between the anchor SNP and each potential partner SNP. The correlations are mapped to a score between 0 and 100, with 100 being the best. The SNP with the highest score is chosen as the partner SNP. The curves below (figure 2) show the correlation-score maps assuming disease models which are multiplicative, dominant and recessive.

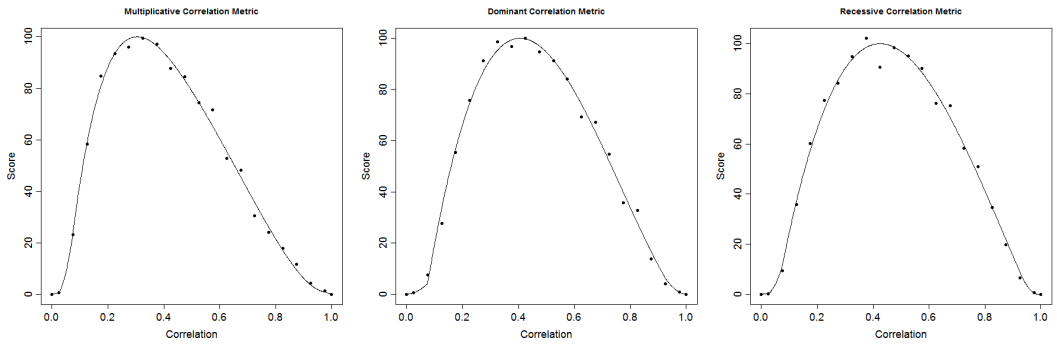

Figure 2: Plots of the partner SNP correlation metrics. Left to right, assuming a multiplicative, dominant and recessive causal variant model respectively.

It is recommended that the default multiplicative causal variant model is used if the penetrances of any causal variants are unknown. The top 10 percent correlations are then in the interval [0.26, 0.36].

If it is known that the causal variant is dominant (recessive) then it should be beneficial to use the dominant (recessive) correlation metric using the `-dominant` (`-recessive`) option. (Note that the AI test still uses a multiplicative model and that this option only affects how the partner SNP is chosen.)

Again, for more details concerning the methodology, please read the accompanying manuscript Howey and Cordell (2014).

6 SNP Window Size

The window size of the SNP window used to pick the partner SNP for each anchor SNP is configurable using the `-window-size` option. It is only possible to set the SNP window size to an even number since the SNP window is in front and behind the anchor SNP. For example, to set the SNP window to size 12:

```
./snipsnip -window-size 12 -o myresults.dat mydata.bed
```

There is no "correct" window size, but 10 has proved in practise to be a good comprise between too small and too big, and is therfore set as the default. It is unlikely that a SNP window size of 20 or greater will be useful. If performing multiple GWASs with different SNP window sizes care should be exercised for multiple testing issues.

It is also possible to set the SNP window using base pair position, so that any SNP within a certain base pair position distance of the anchor SNP is considered for the partner SNP. For example, to set the SNP window a base pair position size of 110,000:

```
./snipsnip -window-size-bp 110000 -o myresults.dat mydata.bed
```

However, it is recommended to use a fixed number of SNPs for the SNP window to increase the chances of picking a good partner SNP.

Again, for more details concerning the methodology, please read the accompanying manuscript Howey and Cordell (2014).

7 Linear Regression

It is possible to use a quantitative trait for the phenotype instead of the case-control status. To do this use the `-linear` option. For example:

```
./snipsnip -linear -o myresults.dat mydata.bed
```

In this case the phenotype column in the `.fam` file should contain the phenotype of interest. The missing quantitative trait value is set by default to -9. This can be changed with the `-mqtv` option. For example, to set the missing quantitative trait value to 0:

```
./snipsnip -linear -mqtv 0 -o myresults.dat mydata.bed
```

When using linear regression the partner SNP is chosen in exactly the same manner as before and the same equivilent models are compared, but this time using an F-test.

Again, for more details concerning the methodology, please read the accompanying manuscript Howey and Cordell (2014).

8 Covariates

It is possible to perform the AI test with a set of covariates. To do this use the `-covar` option. For example:

```
./snipsnip -covar covariates.dat -o myresults.dat mydata.bed
```

The format of the covariate file is the same as PLINK covariate files. That is, a text file where the first column is the pedigree ID, the second column is the individual ID and the remaining columns are the covariate values, where a value of -9 denotes a missing value (this may be changed with the `-mqtv` option). For example, a covariate file with 3 covariates may look as follows:

```
PEDID ID SMOKE ALCOHOL EX
WXA_T1233 WXA_T120 0.0037 0.0033 0.0207
WXA_T1233 WXA_T121 -0.0019 0.022 0.0257
WXA_T1234 WXA_T987 0.0104 0.0096 -0.0154
...
```

The header line may be present or not. The covariates may be chosen with the header names as follows:

```
./snipsnip -covar covariates.dat -covar-name ALCOHOL,EX -o myresults.dat mydata.bed
```

or

```
./snipsnip -covar covariates.dat -covar-name ALCOHOL-EX -o myresults.dat mydata.bed
```

to include all covariates between and including these two. Note that no spaces should be used between the chosen covariate values. The covariates may also be chosen by their numbers. So the above may be written:

```
./snipsnip -covar covariates.dat -covar-number 2,3 -o myresults.dat mydata.bed
```

or

```
./snipsnip -covar covariates.dat -covar-number 2-3 -o myresults.dat mydata.bed
```

9 SnipSnip Examples

9.1 Basic example

Using the example data given with the SnipSnip download, perform a basic SnipSnip analysis as follows:

```
./snipsnip -o results-ExampleData.dat exampleData.bed
```

In R type:

```
exampleData<-read.table("results-ExampleData.dat", header=T)

plot(exampleData$BP/10^6, -log10(exampleData$P), main="SnipSnip Test",
 xlab=expression(bp~position~(Mb)), ylab=expression(-log[10](p-value)), ylim=c(0,15))
abline(h=8, lty=2)
```

This will produce the following plot:
To also perform standard logistic regression use the `-lr` option as follows:

```
./snipsnip -lr -o results-ExampleData-LR.dat exampleData.bed
```

To plot the standard logistic regression test results type:

```
exampleDataLR<-read.table("results-ExampleData-LR.dat", header=T)

plot(exampleDataLR$BP/10^6, -log10(exampleDataLR$P_LR), main="Standard Logistic Regre
 xlab=expression(bp~position~(Mb)), ylab=expression(-log[10](p-value)), ylim=c(0,15))
abline(h=8, lty=2)
```

This will produce the following plot:

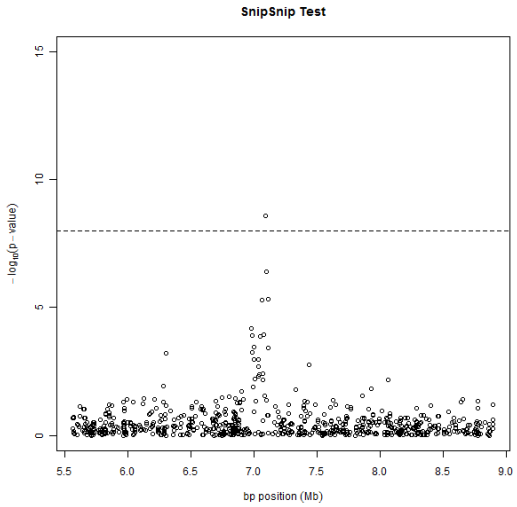

Figure 3: Manhattan plot of AI test results.

9.2 X chromosome

SnipSnip does not do anything special with the analysis of SNPs on the X chromosome - the SNPs will be treated as if they were autosomal. (So females will end up with genotypes coded 0,1,2 and males will end up with genotypes coded 0,2). This has the potential to create false positives if you are dealing with a disease that has different prevalance in males and females; for this reason it is recommended to always include gender as a covariate if you are analysing X-chromosomal SNPs. See section 8 for details on handling covariates.

If you have no other covariates to consider then it is possible to use the `.fam` file as the covariate file by choosing the sex column as the covariate as follows:

```
./snipsnip -covar chromosome23.fam -covar-number 3 -o myresults-chr23.dat chromosome2
```

References

Chen Y, Zhou J, Cheng Z, Yang S, Chu H, Fan Y, Li C, Ho-Yin Wong B, Zheng S, Zhu Y, Yu F, Wang Y, Liu X, Gao H, Yu L, Tang L, Cui D, Hao K, Boss Y, Obeidat M, Brandsma C, Song Y, Kai-Wang To K, Chung Sham P, Yuen K, Li L. 2015. Functional variants regulating LGALS1 (Galectin 1) expression affect human susceptibility to influenza A(H7N9). Scientific Reports 5:8517.

Howey R, Cordell HJ. 2014. Imputation without doing imputation: a new method for the detection of non-genotyped causal variants. Genet Epidemiol 38:173–190.

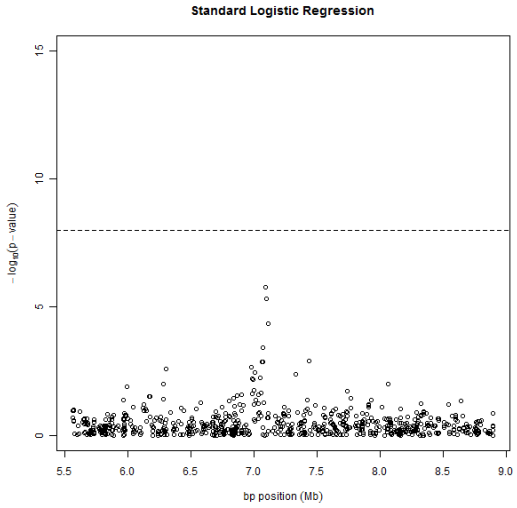

Figure 4: Manhattan plot of standard logistic regression test results.