ASPBay package vignette
Version 1.1

Claire DANDINE-ROULLAND

January 14, 2015

1 Introduction

In the Package `ASPBay`, we propose an algorithm in two steps using controls and affected sib-pairs data to make inference on causal variants (for more details, DANDINE-ROULLAND, Claire and PERDRY, Herve, *Where is the causal variant? On the advantage of the family design over the case-control design in genetic association studies*, Submitted to Eur J Hum Genet).

The first step is to select a subset of SNPs which likely contains the causal SNPs or SNPs in linkage disequilibrium with them. For this, we compute an association statistic for each SNP:

$$Y = U/\sqrt{\widehat{\sigma^2}}$$

where U is the score

$$U = \left(\sum_{k,i \in \{0,1,2\}} (2+i) n_{ki} \right) \hat{f} + \frac{1}{2} \sum_{k,i \in \{0,1,2\}} (2+i) k n_{ki}$$

with n_{ki} the number of ASPs in which the index genotype is k and the number of IBD alleles is i, m_k the number of controls with genotype k, n and m the total number of affected sib-pairs and controls, and $\hat{f} = \frac{\left(\sum_{i \in \{0,1,2\}} n_{1i} + 2 \sum_{i \in \{0,1,2\}} n_{2i} + m_1 + 2m_2\right)}{2(m+n)}$. And

$$\widehat{\sigma^2} = \frac{1}{4} \times \frac{\left(1 - \hat{f}\right) \hat{f} \left(19m + n - 1\right) n}{n+m}$$

the estimator of the variance of U under the hypothesis of no association.

We keep the SNPs j which verify

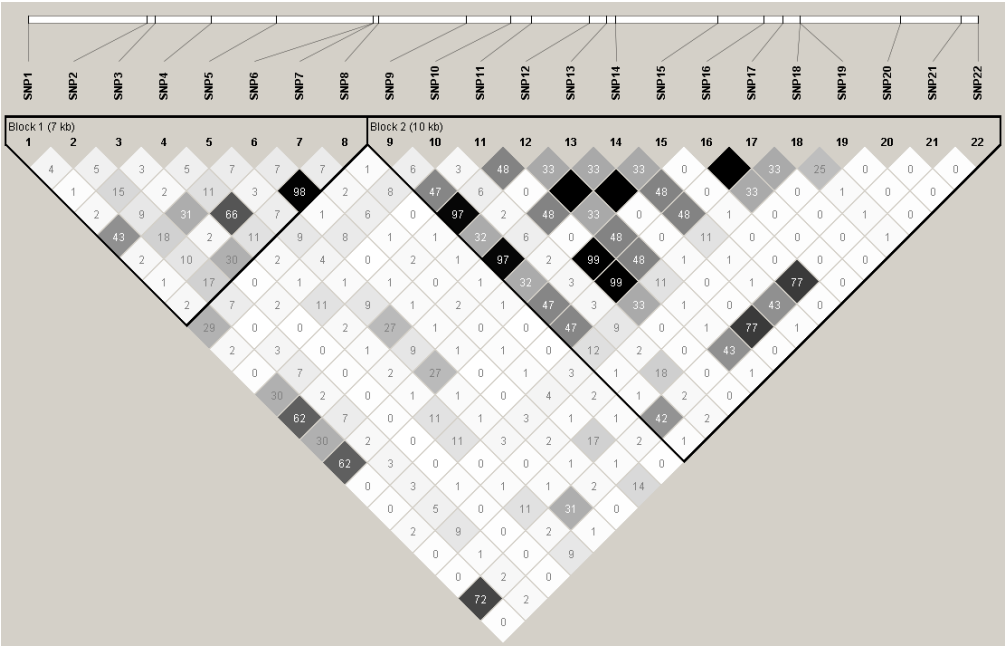
$$\max_l(Y_l^2) - Y_j^2 < \mathrm{k}$$

with k the chosen threshold.

In the second step, we use the SNPs selected by the first step to retrieve information on the causal variants of the region. Let's assume that a variant B in linkage disequilibrium with the causal variant A is observed. In this case, we want to make inferences on A, in particular to estimate the LD between A and B, and the OR of A. For that, we sample in the posterior distribution of frequencies parameters, linkage disequilibrium and causal odds ratio (OR) using Metropolis-Hastings algorithm.

2 An example with a simulated dataset

The package provides an simulated dataset named `ASPData`. We simulated 1000 controls and affected sib pairs. The simulated genomic region consists in 22 genotypes along them only one is causal. The odds ratio of the causal SNP is 2 and its frequency of alternative allele is 0.0754. For the linkage disequilibrium, we use the structure of European population.



The dataset `ASPData` consists in a list with components

- `Control`, the matrix of the 21 genotypes of controls (causal SNP not included)
- `Index`, the matrix of the 21 genotypes of index cases (causal SNP not included)
- `IBD`, the vector of the IBD states for each affected sib pairs
- `Causal`, the name of the causal SNP

```
> data(ASPData)
```

First, we test the association of each observed SNPs with score test.

```
> Score <- ASP.Score(ASPData$Control, ASPData$Index, ASPData$IBD)
> Score$Value
```

```
      SNP1       SNP2       SNP3       SNP4       SNP5       SNP6       SNP7       SNP8
10.5697498 -2.6353276 -0.7334053 -0.7617181  6.3666141 -2.4239793 -0.3127143 -2.5229966
      SNP9      SNP10      SNP11      SNP12      SNP13      SNP14      SNP15      SNP16
 9.2774756 -3.7236812 -0.9274910  9.4499055 11.2370354  9.4499055 11.2370354 -0.9820782
     SNP17      SNP18      SNP19      SNP20      SNP22
-0.9820782 -1.5493967 -0.4868593 -0.1776845 -0.6855409
```

```
> Score$Pvalue
```

```
        SNP1         SNP2         SNP3         SNP4         SNP5         SNP6
0.000000e+00 8.405611e-03 4.633113e-01 4.462283e-01 1.932465e-10 1.535148e-02
        SNP7         SNP8         SNP9        SNP10        SNP11        SNP12
7.544978e-01 1.163595e-02 0.000000e+00 1.963389e-04 3.536717e-01 0.000000e+00
       SNP13        SNP14        SNP15        SNP16        SNP17        SNP18
0.000000e+00 0.000000e+00 0.000000e+00 3.260613e-01 3.260613e-01 1.212864e-01
       SNP19        SNP20        SNP22
6.263581e-01 8.589708e-01 4.930027e-01
```

We see that, under Bonferroni correction, the SNPs 1, 9, 10, 12, 13, 14, 15 have an significative association test.

Now, we want to select in this SNPs a subset to capture information about the true causal SNP.

```
> Select <- ASP.Selection(ASPData$Control, ASPData$Index, ASPData$IBD)
> Select$SNPnames_subset
```

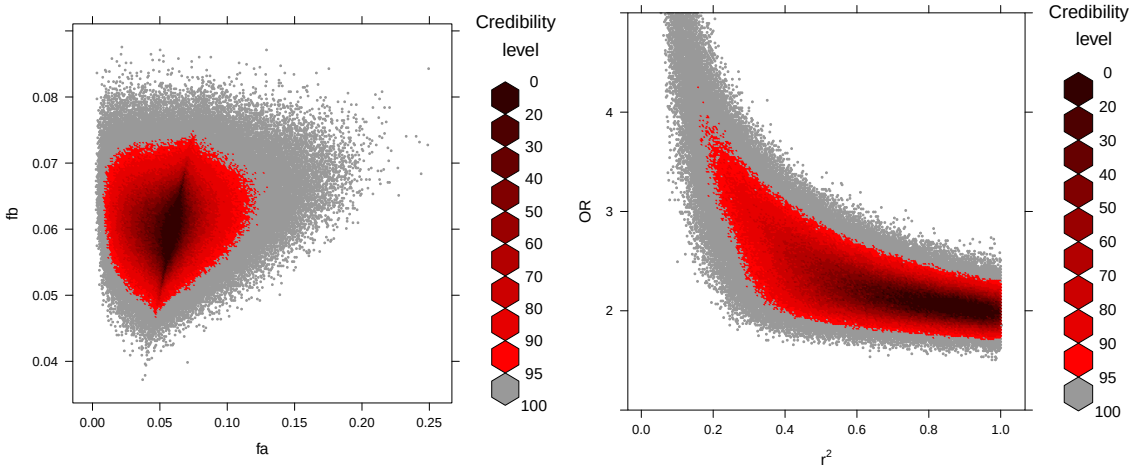
```
[1] "SNP13" "SNP15"
```

The algorithm select two SNPs, the SNPs 13 and 15. These two SNPs are in complete linkage disequilibrium, hence we apply the Metropolis-Hasting algorithm only for the SNP 15. This SNP have an alternative allele frequency of 0.059 and a linkage disequilibrium with the causal SNP of 0.77.

```
> M15 <- ASP.Bayesian(1e7, ASPData$Control, ASPData$Index, ASPData$IBD, 15, thin = 10, sd.psi=0.03)
> G15 <- Graphs.Bayesian(M15, burn = 1000, print=FALSE)
```

```
> print(G15$hex_fa_fb)
```

```
> print(G15$hex_r2_OR)
```

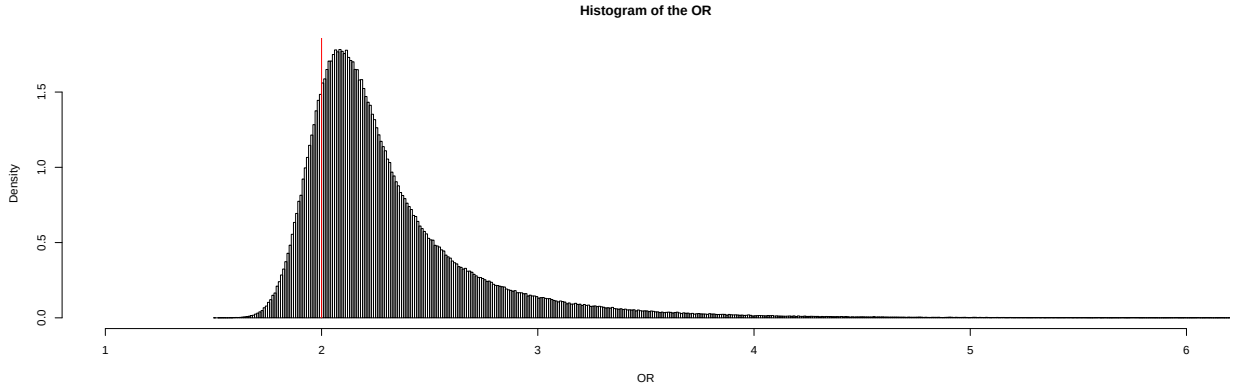


Here, we see that the true values of parameters are in the credibility regions.

We can also make the histograms of parameters to see their marginal distributions.

We begin by the histogram of the causal OR. Before, the causal odds ratio is transformed. The value of OR is kept if it is superior to 1 or it is inversed if it is inferior to 1. This step avoids to obtain two peaks corresponding to equivalent parameter values.

```
> OR <- M15$OR*(M15$OR>=1) + 1/M15$OR*(M15$OR<1)
> hist(OR[-(1:10000)], freq=FALSE, breaks=1000, main='Histogram of the OR',
+      xlab='OR', xlim=c(1,6))
> lines(c(2,2), c(0,100) ,type='l', col='red')
```

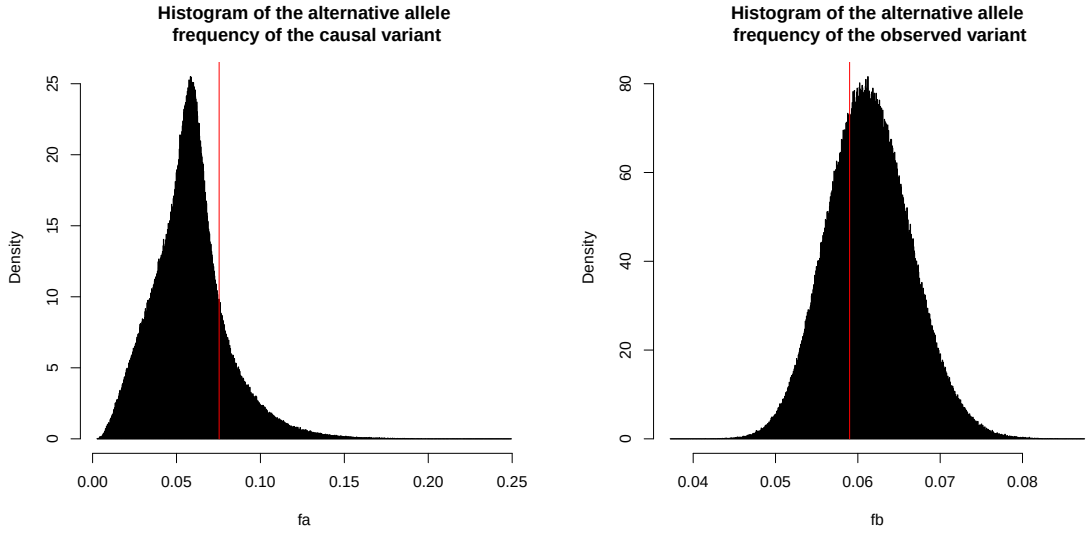


The red line represents the true causal odds ratio. The figure shows that the mod of posterior distribution is very near to the true odds ratio.

We can also consider the frequencies. In the same idea using for the odds ratio, the causal alternative allele frequency is transformed to correspond to an odds ratio superior to 1. The frequency is replaced by its complement to 1 if the OR is inferior to 1.

```
> M15$fa <- M15$f_ab + M15$f_aB
> fa <- M15$fa*(M15$OR>=1) + (1-M15$fa)*(M15$OR<1)
> hist(fa[-(1:10000)], freq=FALSE, breaks=1000,
+             main='Histogram of the alternative allele\n frequency of the causal variant',
+             xlab='fa')
> lines(c(0.0754,0.0754), c(0,100) ,type='l', col='red')
```

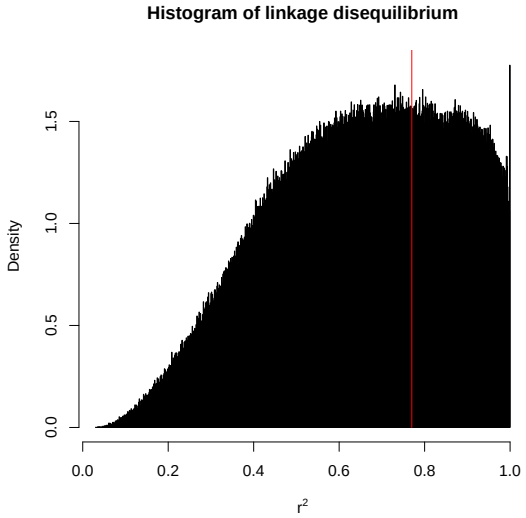
```
> M15$fb <- M15$f_ab + M15$f_Ab
> hist(M15$fb[-(1:10000)], freq=FALSE, breaks=1000,
+             main='Histogram of the alternative allele\n frequency of the observed variant',
+             xlab='fb')
> lines(c(0.059,0.059), c(0,100) ,type='l', col='red')
```



We can see that the distribution of the causal variant frequency is more dispersed that the observed variant frequency. We remark also that the mods are near to the true values of alternative allele frequencies in particular for the observed SNP.

Finally, we see the histogram of the linkage disequilibrium r^2.

```
> M15$D <- M15$f_ab*M15$f_AB - M15$f_aB*M15$f_Ab
> M15$r2 <- M15$D**2/( M15$fa*(1-M15$fa)*M15$fb*(1-M15$fb) )
> hist(M15$r2[-(1:10000)], freq=FALSE, breaks=1000,
+      main='Histogram of linkage disequilibrium', xlab=expression(r^2))
> lines(c(0.77,0.77), c(0,100) ,type='l', col='red')
```



We can see that the distribution is very dispersed but the mod is around to the true value of linkage disequilibrium.