1 Introduction

In genetics association studies, investigators are interested in testing the association between disease and DSL (disease susceptibility locus) in a case-control study. However, DSL is usually unknown. Hence what one can do is to test the association between disease and DSL indirectly by testing the association between disease and a marker. The power of the association between disease and DSL depends on the power of the association between disease and the marker, and on the LD (linkage disequilibrium) between disease and the marker.

Assume that both DSL and the marker are biallelic and assume additive model. Given the MAFs (minor allele frequencies) for both DSL and the marker, relative risks, \( r^2 \) or \( D' \) between DSL and marker, the numbers of cases and controls, and the significant level for hypothesis testing, the tools in the GeneticsDesign package can calculate the power for the association test that disease is associated with DSL in a case-control study. This information can help investigator to determine appropriate sample sizes in experiment design stage.

2 Examples

To call the functions in the R package GeneticsDesign, we first need to load it into R:

> library(GeneticsDesign)

The following is a sample code to get a table of power for different combinations of high risk allele frequency \( Pr(A) \) and genotype relative risk \( RR(Aa|aa) \).
> set1 <- seq(from = 0.1, to = 0.5, by = 0.1)
> set2 <- c(1.25, 1.5, 1.75, 2)
> len1 <- length(set1)
> len2 <- length(set2)
> mat <- matrix(0, nrow = len1, ncol = len2)
> rownames(mat) <- paste("MAF=", set1, sep = "")
> colnames(mat) <- paste("RRAA=", set2, sep = "")
> for (i in 1:len1) {
+    a <- set1[i]
+    for (j in 1:len2) {
+        b <- set2[j]
+        res <- GPC.default(pA = a, pD = 0.1, RRAa = (1 + b)/2,
+            RRAA = b, Dprime = 1, pB = a, quiet = T)
+        mat[i, j] <- res$power
+    }
+}
> print(round(mat, 3))

<table>
<thead>
<tr>
<th>RRAA=1.25</th>
<th>RRAA=1.5</th>
<th>RRAA=1.75</th>
<th>RRAA=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF=0.1</td>
<td>0.145</td>
<td>0.400</td>
<td>0.690</td>
</tr>
<tr>
<td>MAF=0.2</td>
<td>0.214</td>
<td>0.594</td>
<td>0.877</td>
</tr>
<tr>
<td>MAF=0.3</td>
<td>0.259</td>
<td>0.682</td>
<td>0.926</td>
</tr>
<tr>
<td>MAF=0.4</td>
<td>0.280</td>
<td>0.711</td>
<td>0.936</td>
</tr>
<tr>
<td>MAF=0.5</td>
<td>0.282</td>
<td>0.702</td>
<td>0.925</td>
</tr>
</tbody>
</table>

A  Technical Details

Suppose that $A$ is the minor allele for the disease locus; $a$ is the common allele for the disease locus; $B$ is the minor allele for the marker locus; and $b$ is the common allele for the marker locus. We use $D$ to denote the disease status “diseased” and use $\bar{D}$ to denote the disease status “not diseased”.

Given the minor allele frequencies $Pr(A)$ and $Pr(B)$ and Linkage disequilibrium (LD) measure $D'$, we can get haplotype frequencies $Pr(AB)$, $Pr(Ab)$, $Pr(aB)$, and $Pr(ab)$:

\[
Pr(AB) = Pr(A)Pr(B) + D \\
Pr(aB) = Pr(a)Pr(B) - D \\
Pr(Ab) = Pr(A)Pr(b) - D \\
Pr(ab) = Pr(a)Pr(b) + D,
\]

where \[ D = D'd_{max} \]
and
\[
d_{\text{max}} = \begin{cases} 
\min \{Pr(A)Pr(b), Pr(a)Pr(B)\}, & \text{if } D > 0, \\
\max \{-Pr(A)Pr(B), -Pr(a)Pr(b)\}, & \text{if } D < 0.
\end{cases}
\]

That is,
\[
D = Pr(AB) - Pr(A)Pr(B).
\]

Note that \( D > 0 \) means \( Pr(AB) > Pr(A)Pr(B) \), i.e., the probability of occurring the haplotype \( AB \) is higher than the probability that the haplotype \( AB \) occurs merely by chance. \( D < 0 \) means \( Pr(AB) < Pr(A)Pr(B) \), i.e., the probability of occurring the haplotype \( AB \) is smaller than the probability that the haplotype \( AB \) occurs merely by chance. In other words, given the same sample size, we can observe haplotype \( AB \) more often when \( D > 0 \) than when \( D < 0 \). Hence, \textbf{we assume that} \( D > 0 \).

\( D \) can also be rewritten as
\[
D = Pr(B|A)Pr(A) - Pr(A)Pr(B) = [Pr(B|A) - Pr(A)] Pr(A).
\]

Hence \( D > 0 \) is equivalent to \( Pr(B|A) > Pr(A) \) which indicates positive association between minor allele \( A \) in disease locus and minor allele \( B \) in marker locus.

The relative risks are
\[
RR(AA|aa) = \frac{Pr(D|AA)}{Pr(D|aa)}
\]
\[
RR(Aa|aa) = \frac{Pr(D|Aa)}{Pr(D|aa)}
\]

The disease prevalence \( Pr(D) \) can be rewritten as
\[
Pr(D) = Pr(D|AA)Pr(AA) + Pr(D|Aa)Pr(Aa) + Pr(D|aa)Pr(aa).
\]

Dividing both sides by \( Pr(D|aa) \), we get
\[
\frac{Pr(D)}{Pr(D|aa)} = \frac{Pr(D|AA)}{Pr(D|aa)} Pr(AA) + \frac{Pr(D|Aa)}{Pr(D|aa)} Pr(Aa) + \frac{Pr(D|aa)}{Pr(D|aa)} Pr(aa)
\]
\[
= RR(AA|aa) [Pr(A)]^2 + RR(Aa|aa)2Pr(A)Pr(a) + [Pr(a)]^2.
\]

Hence
\[
Pr(D|aa) = \frac{Pr(D)}{RR(AA|aa)[Pr(A)]^2 + RR(Aa|aa)2Pr(A)Pr(a) + [Pr(a)]^2}
\]
\[
Pr(D|Aa) = RR(Aa|aa)Pr(D|aa)
\]
\[
Pr(D|AA) = RR(AA|aa)Pr(D|aa)
\]

3
Once we obtain the penetrances of disease locus ($Pr(D|aa)$, $Pr(D|Aa)$, and $Pr(D|AA)$), we can get the sampling probabilities:

\[
Pr(BB|D) = \frac{Pr(D, BB)}{Pr(D)} = \frac{Pr(D, BB, AA) + Pr(D, BB, Aa) + Pr(D, BB, aa)}{Pr(D)}
\]

\[
= \frac{Pr(D|BB, AA)Pr(BB, AA) + Pr(D|BB, Aa)Pr(BB, Aa) + Pr(D|BB, aa)Pr(BB, aa)}{Pr(D)}
\]

\[
Pr(Bb|D) = \frac{Pr(D, Bb)}{Pr(D)} = \frac{Pr(D, Bb, AA) + Pr(D, Bb, Aa) + Pr(D, Bb, aa)}{Pr(D)}
\]

\[
= \frac{Pr(D|Bb, AA)Pr(Bb, AA) + Pr(D|Bb, Aa)Pr(Bb, Aa) + Pr(D|Bb, aa)Pr(Bb, aa)}{Pr(D)}
\]

\[
Pr(bb|D) = \frac{Pr(D, bb)}{Pr(D)} = \frac{Pr(D, bb, AA) + Pr(D, bb, Aa) + Pr(D, bb, aa)}{Pr(D)}
\]

\[
= \frac{Pr(D|bb, AA)Pr(bb, AA) + Pr(D|bb, Aa)Pr(bb, Aa) + Pr(D|bb, aa)Pr(bb, aa)}{Pr(D)}
\]

We assume that

\[Pr(D|BB, AA) = Pr(D|AA), \quad Pr(D|BB, Aa) = Pr(D|Aa), \quad Pr(D|BB, aa) = Pr(D|aa)\]

We also have

\[Pr(BB, AA) = [Pr(AB)]^2\]
\[Pr(BB, Aa) = 2Pr(AB)Pr(aB)\]
\[Pr(BB, aa) = [Pr(ab)]^2\]
\[Pr(Bb, AA) = 2Pr(AB)Pr(Ab)\]
\[Pr(Bb, Aa) = 2[Pr(AB) \times Pr(ab) + Pr(AB)Pr(ab)]\]
\[Pr(Bb, aa) = 2Pr(aB)Pr(ab)\]

\[Pr(bb, AA) = [Pr(AB)]^2\]
\[Pr(bb, Aa) = 2Pr(AB)Pr(ab)\]
\[Pr(bb, aa) = [Pr(ab)]^2\]
Hence

\[
\Pr(BB|D) = \frac{\Pr(D|BB, AA)\Pr(BB, AA) + \Pr(D|BB, Aa)\Pr(BB, Aa) + \Pr(D|BB, aa)\Pr(BB, aa)}{\Pr(D)}
\]

\[
= \frac{\Pr(D|AA) [\Pr(AB)]^2 + \Pr(D|Aa) [2\Pr(AB)\Pr(ab)] + \Pr(D|aa) [\Pr(ab)]^2}{\Pr(D)}
\]

\[
\Pr(Bb|D) = \frac{\Pr(D|Bb, AA)\Pr(Bb, AA) + \Pr(D|Bb, Aa)\Pr(Bb, Aa) + \Pr(D|Bb, aa)\Pr(Bb, aa)}{\Pr(D)}
\]

\[
= \frac{\Pr(D|AA) [2\Pr(AB)\Pr(ab)] + \Pr(D|Aa) \{2 [\Pr(AB)\Pr(ab) + \Pr(AB)\Pr(ab)]\}}{\Pr(D)}
\]

\[
+ \frac{\Pr(D|aa) [2\Pr(ab)\Pr(ab)]}{\Pr(D)}
\]

\[
\Pr(bb|D) = \frac{\Pr(D|bb, AA)\Pr(bb, AA) + \Pr(D|bb, Aa)\Pr(bb, Aa) + \Pr(D|bb, aa)\Pr(bb, aa)}{\Pr(D)}
\]

\[
= \frac{\Pr(D|AA) [\Pr(ab)]^2 + \Pr(D|Aa) [2\Pr(ab)\Pr(ab)] + \Pr(D|aa) [\Pr(ab)]^2}{\Pr(D)}
\]

To calculate the sampling probabilities \(\Pr(BB|\bar{D})\), \(\Pr(Bb|\bar{D})\), and \(\Pr(bb|\bar{D})\), we can apply Bayesian rules again:

\[
\Pr(BB|\bar{D}) = \frac{\Pr(\bar{D}|BB)\Pr(BB)}{\Pr(\bar{D})} = \frac{[1 - \Pr(D|BB)][\Pr(B)]^2}{1 - \Pr(D)}
\]

\[
\Pr(Bb|\bar{D}) = \frac{\Pr(\bar{D}|Bb)\Pr(Bb)}{\Pr(\bar{D})} = \frac{[1 - \Pr(D|Bb)][2\Pr(B)\Pr(b)]}{1 - \Pr(D)}
\]

\[
\Pr(bb|\bar{D}) = \frac{\Pr(\bar{D}|bb)\Pr(bb)}{\Pr(\bar{D})} = \frac{[1 - \Pr(D|bb)][\Pr(b)]^2}{1 - \Pr(D)}
\]

and

\[
\Pr(D|BB) = \frac{\Pr(BB|D)\Pr(D)}{\Pr(BB)} = \frac{\Pr(BB|D)\Pr(D)}{[\Pr(B)]^2}
\]

\[
\Pr(D|Bb) = \frac{\Pr(Bb|D)\Pr(D)}{\Pr(Bb)} = \frac{\Pr(Bb|D)\Pr(D)}{[2\Pr(B)\Pr(b)]}
\]

\[
\Pr(D|bb) = \frac{\Pr(bb|D)\Pr(D)}{\Pr(bb)} = \frac{\Pr(bb|D)\Pr(D)}{[\Pr(b)]^2}
\]
The expected allele frequencies are

\[ Pr(B|D) = \frac{2n_{case} Pr(BB|D)}{2n_{case}} + \frac{n_{case} Pr(Bb|D)}{2n_{case}} = Pr(BB|D) + Pr(Bb|D)/2 \]

\[ Pr(b|D) = \frac{2n_{case} Pr(bb|D)}{2n_{case}} + \frac{n_{case} Pr(Bb|D)}{2n_{case}} = Pr(bb|D) + Pr(Bb|D)/2 \]

\[ Pr(B|\bar{D}) = Pr(BB|\bar{D}) + Pr(Bb|\bar{D})/2 \]
\[ Pr(b|\bar{D}) = Pr(bb|\bar{D}) + Pr(Bb|\bar{D})/2. \]

The counts for cases

\[ n_{2c} = n_{cases} Pr(BB|D), \quad n_{1c} = n_{cases} Pr(Bb|D), \quad n_{0c} = n_{cases} Pr(bb|D). \]

The counts for controls

\[ n_{2n} = n_{controls} Pr(BB|\bar{D}), \quad n_{1n} = n_{controls} Pr(Bb|\bar{D}), \quad n_{0n} = n_{controls} Pr(bb|\bar{D}). \]

Table 1: Counts for cases and controls

<table>
<thead>
<tr>
<th>marker genotype</th>
<th>cases</th>
<th>controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>(n_{2c})</td>
<td>(n_{2n})</td>
</tr>
<tr>
<td>Bb</td>
<td>(n_{1c})</td>
<td>(n_{1n})</td>
</tr>
<tr>
<td>bb</td>
<td>(n_{0c})</td>
<td>(n_{0n})</td>
</tr>
<tr>
<td>total</td>
<td>(n_{cases})</td>
<td>(n_{controls})</td>
</tr>
</tbody>
</table>

Odds ratios are

\[ OR(BB|bb) = \frac{n_{2c}n_{0n}}{n_{2n}n_{0c}} = \frac{Pr(BB|D) Pr(bb|\bar{D})}{Pr(Bb|D) Pr(b|\bar{D})} \]
\[ OR(Bb|bb) = \frac{n_{1c}n_{0n}}{n_{1n}n_{0c}} = \frac{Pr(Bb|D) Pr(bb|\bar{D})}{Pr(Bb|D) Pr(b|\bar{D})} \]

Define

\[ U = \sum_{i=1}^{3} x_i \left( \frac{S}{N} r_i - \frac{R}{N} s_i \right), \]

where

\[ R = n_{0c} + n_{1c} + n_{2c}, \]
\[ S = n_{0n} + n_{1n} + n_{2n}, \]
\[ N = R + S. \]
We have
\[ V = \overline{\text{Var}}(U) = \frac{RS}{N^3} \left[ N \sum_{i=1}^{3} x_i^2 n_i - \left( \sum_{i=1}^{3} x_i n_i \right)^2 \right]. \]

The Linear Trend Test Statistic
\[ Z_T^2 = \left( \frac{U}{\sqrt{V}} \right)^2 \]
asymptotically follows a \( \chi^2_1 \) distribution under the null hypothesis. Under alternative hypothesis, \( Z_T^2 \) asymptotically follows a non-central chi-square distribution \( \chi^2_{1, \lambda_T} \) with non-centrality parameter
\[
\lambda_T = RS \frac{\left[ \sum_{i=1}^{3} x_i(p_{1i} - p_{0i}) \right]^2}{\sum_{i=1}^{3} x_i^2 (Rp_{0i} + Sp_{1i}) - \left[ \sum_{i=1}^{3} x_i(Rp_{0i} + Sp_{1i}) \right]^2 / N},
\]
where
\[
x_1 = 2, \ x_2 = 1, \ x_3 = 0,
\]
\[
n_1 = n_{2c} + n_{2n}, \ n_2 = n_{1c} + n_{1n}, \ n_3 = n_{0c} + n_{0n},
\]
\[
p_{01} = Pr(BB|\bar{D}), \ p_{02} = Pr(Bb|\bar{D}), \ p_{03} = Pr(bb|\bar{D}),
\]
\[
p_{11} = Pr(BB|D), \ p_{12} = Pr(Bb|D), \ p_{13} = Pr(bb|D).
\]

Given significant level \( \alpha \), the power \( 1 - \beta \) satisfies:
\[
\text{Pr}(Z_T^2 > c|H_0) = \alpha
\]
\[
\text{Pr}(Z_T^2 > c|H_a) = 1 - \beta.
\]

References


