Published in final edited form as:
Methods Mol Biol. 2012 ; 850: . doi:10.1007/978-1-61779-555-8_18.

Single Marker Association Analysis for Unrelated Samples

Gang Zheng, **Jinfeng Xu**, **Ao Yuan**, and **Joseph L. Gastwirth**

Abstract

Methods for single marker association analysis are presented for binary and quantitative traits. For a binary trait, we focus on the analysis of retrospective case–control data using Pearson's chi-squared test, the trend test, and a robust test. For a continuous trait, typical methods are based on a linear regression model or the analysis of variance. We illustrate how these tests can be applied using a public available R package "Rassoc" and some existing R functions. Guidelines for choosing these test statistics are provided.

Keywords

Additive; Association; ANOVA; Binary trait; Case–control design; Dominant; Genetic model; Genotype relative risks; MAX3; Mode of inheritance; Penetrance; Rassoc; Recessive; Quantitative trait; Robustness

1. Introduction

Statistical procedures for testing whether there is an association between a phenotype and a single nucleotide polymorphism (SNP) are described and illustrated. Usually, the phenotype of interest is either a binary or quantitative one. For a binary trait, we focus on a retrospective case–control study, in which cases and controls are randomly drawn from case and control populations, respectively. For a continuous trait, the data are obtained from a random sample of the general population. Although a large number of SNPs is available for testing association, single marker analysis is often employed. The significance level to test a single hypothesis is 0.05. When multiple SNPs are tested, the Bonferroni correction can be applied.

Denote the genotypes of an SNP as G_0, G_1, and G_2. For case–control data, denote the penetrances as f_0, f_1, and f_2 with respect to the three genotypes, respectively. Under the null hypothesis H_0, we have $f_0 = f_1 = f_2 = \Pr(\text{case})$. A genetic model is recessive if $f_1 = f_0$, additive if $f_1 = (f_0 + f_2)/2$, or dominant if $f_1 = f_2$. For a single SNP, the observed case–control data consist of genotype counts (r_0, r_1, r_2) among r cases and (s_0, s_1, s_2) among s controls. Denote $n_j = r_j + s_j$ ($j = 0, 1, 2$) and $n = r + s$. The Cochran-Armitage trend test (referred to as the trend test) is one of the two most commonly used statistics for the analysis of case–control data. It can be written as (1, 2)

$$T_1(x) = \frac{\sum_{j=0}^{2} x_j \{(1-\varphi) r_j - \varphi s_j\}}{\left[n\varphi(1-\varphi) \left\{ \sum_{j=0}^{2} x_j^2 n_j/n - \left(\sum_{j=0}^{2} x_j n_j/n \right)^2 \right\} \right]^{1/2}} \quad (1)$$

where $(x_0, x_1, x_2) = (0, x, 1)$, x is determined by the genetic model and $\phi = r/n$. Under H_0, given x, $T_1(x)$ asymptotically follows a standard normal distribution $N(0,1)$. When the genetic model is recessive and the risk allele is known, $T_1(0)$ is used. When the genetic model is dominant and the risk allele is known, $T_1(1)$ is used. When we only know that the genetic model is recessive (or dominant) but not the risk allele, $T_1(0)$ (or $T_1(1)$) cannot be used alone. When the genetic model is additive regardless of the risk allele, $T_1(1/2)$ is used. Pearson's chi-squared test (referred to as Pearson's test) is another commonly used test. It can be written as

$$T_2 = \sum_{j=0}^{2} \frac{(r_j - rn_j/n)^2}{(rn_j/n)} + \sum_{j=0}^{2} \frac{(s_j - sn_j/n)^2}{(sn_j/n)}.$$

Under H_0, T_2 asymptotically follows a chi-squared distribution with two degrees of freedom (d*f*), denoted as χ^2_2. The robust test, MAX3, is given by

$$\mathrm{MAX3} = \max\left(|T_1(0)|, |T_1(1/2)|, T_1(1)|\right),$$

where $T_1(0)$, $T_1(1/2)$, and $T_1(1)$ are the trend tests given by Eq. 1 with different x values. Under H_0, the asymptotic distribution of MAX3 is far more complex than those of $T_1(x)$ and T_2. A procedure for determining its asymptotic null distribution and *P*-value is discussed in Note 2. For other robust tests, see Note 3.

The power of each statistical test depends on the underlying genetic model. It will be seen that MAX3 is more robust than any single trend test or Pearson's chi-squared test when the genetic model is unknown, thus, it should be used in practice. Since MAX3 does not follow any chi-squared distribution, we discuss how to find its *P*-value using the R package Rassoc (3). This package is available from the Comprehensive R Archive Network at "http://CRAN.R-project.org/package=Rassoc."

For a continuous trait Υ, a typical model is $\mathrm{Y} = \mu + g + \in$, where μ is a fixed overall mean of the trait under H_0, g is the random genetic effect due to G, and $\in$ is a random error. The genetic value of g is $-a$ when $G = G_0$, d when $G = G_1$, and a when $G = G_2$. Under H_0, we have $a = d = 0$. A genetic model is recessive, additive, or dominant if $d = a$, $d = 0$, or $d = a$, respectively. The observed data consist of pairs (Y_{ij}, G_j) for $i = 1,\ldots, n_j$ and $j = 0, 1, 2$, where Y_{ij} is the trait value of the ith individual with genotype G_j. Denote $n = n_0 + n_1 + n_2$.

For the analysis of a quantitative trait, linear regression and the analysis of variance (ANOVA) are routinely used. Let $(x_0, x_1, x_2) = (0, x, 1)$ be the values for the genotypes (G_0, G_1, G_2), where $x = 0$, 1/2, or 1 for the recessive, additive, or dominant models, respectively. Using the data $(\mathrm{Y}_{ij}, G_i), i = 1, \ldots, n_j$ and $j = 0, 1, 2$, the *F*-test derived from a linear regression model is given by

$$F(x) = \frac{(n-2)\left\{\sum_j \sum_i \left(x_j - \bar{x}\right)\left(\mathrm{Y}_{ij} - \bar{\mathrm{Y}}\right)^2\right\}}{\sum_j \sum_i \left(x_j - \bar{x}\right)^2 \sum_j \sum_i \left(\mathrm{Y}_{ij} - \bar{\mathrm{Y}}\right)^2 - \left\{\sum_j \sum_i \left(x_j - \bar{x}\right)\left(\mathrm{Y}_{ij} - \bar{\mathrm{Y}}\right)\right\}^2}, \quad (2)$$

where $\bar{x} = \sum_{j=0}^{2} n_j x_j / n$ and $\bar{\mathrm{Y}} = \sum_{j=0}^{2} \sum_{i=1}^{n_j} \mathrm{Y}_{ij} / n$. Given x, $F(x)$ has an asymptotic *F*-distribution with $(1, n-2)$ d*f* under H_0.

Alternatively, denote $\bar{Y}_{.j}=\sum_{i=1}^{n_j} Y_{ij}/n_j$ for j = 0, 1, 2, $SS_b=\sum_{j=0}^{2} n_j\left(\bar{Y}_{.j}-\bar{Y}\right)^2$ and $SS_e=\sum_{j=0}^{2}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_{.j}\right)^2$. The *F*-test derived from the ANOVA is given by

[2]The asymptotic null distribution and *P*-value of MAX3.

Three approaches to compute the *P*-value of MAX3 are presented in ref. 3. Let ρ_{xy} be the asymptotic null correlation of $T_1(x)$ and $T_1(y)$, where x,y = 0, 1/2, 1, and let p_j be the population frequency of genotype G_j (j = 0, 1, 2). Then

$$\rho_{xy}=\frac{(xyp_1+p_2)-(xp_1+p_2)(yp_1+p_2)}{\left\{(x^2p_1+p_2)-(xp_1+p_2)^2\right\}^{1/2}\left\{(y^2p_1+p_2)-(yp_1+p_2)^2\right\}^{1/2}}.$$

Denote $\omega_0=\left(\rho_{0\frac{1}{2}}-\rho_{01}\rho_{\frac{1}{2}1}\right)/\left(1-\rho_{01}^2\right)$ and $\omega_1=\left(\rho_{\frac{1}{2}1}-\rho_{01}\rho_{0\frac{1}{2}}\right)/\left(1-\rho_{01}^2\right)$. In the following, ρ_{xy}, ω_0, and ω_1 are estimated under H_0 by replacing p_j with $\hat{p_j}=n_j/n$ (j = 1,2).

The asymptotic distribution of MAX3 under H_0 is far more complex than those of the trend test and Pearson's test. An expression for the asymptotic null distribution of MAX3, $P(t)$ = Pr(MAX3 < t), is given by

$$\begin{aligned}P(t)=&2\int_0^{\frac{t(1-\omega_1)}{\omega_0}}\Phi\left(\frac{t-\rho_{01}u}{\sqrt{1-\rho_{01}^2}}\right)\phi(u)\quad \mathrm{d}u,\\&-2\int_0^{t}\Phi\left(\frac{-t-\rho_{01}u}{\sqrt{1-\rho_{01}^2}}\right)\phi(u)\quad \mathrm{d}u+2\int_{\frac{t(1-\omega_1)}{\omega_0}}^{t}\Phi\left(\frac{t-\omega_0u/\omega_1-\rho_{01}u}{\sqrt{1-\rho_{01}^2}}\right)\phi(u)\quad \mathrm{d}u,\end{aligned}\tag{4}$$

where φ and $\mathbf{\Phi}$ are the density and distribution functions of N (0,1) (3). Using Eq. 4, the asymptotic *P*-value of MAX3 is given by 1 – *P*(max3), where max3 is the observed MAX3. This approach is denoted as "asy" in the second R function in Rassoc.

Alternatively, simulations can be used to approximate the null distribution of MAX3. Given the observed data (r_0, r_1, r_2) and (s_0, s_1, s_2), in the jth simulation (j = 1,. . ., m), we generate (r_{0j}, r_{1j}, r_{2j}) from the multinomial distribution Mul(r; $\hat{p_0}$, $\hat{p_1}$, $\hat{p_2}$) and (s_{0j}, s_{1j}, s_{2j}) from the same distribution except that r is replaced by s, where $\hat{p_i}=n_i/n$ (i = 0, 1, 2). For each j, we compute MAX3 denoted as MAX3j. Then MAX3$_1$, . . ., MAX3$_m$ form an empirical null distribution of MAX3 when m is large enough. For single marker analysis, we use m = 100,000 to determine the null distribution. A larger m may be used, if the *P*-value of MAX3 is smaller than 10^{-5}. This parametric bootstrap procedure is denoted as "boot" in the second R function in Rassoc.

A more efficient simulation approach, denoted as "bvn" in the second R function in Rassoc, is to directly generate $T_1(0)$ and $T_1(1)$ in the jth simulation from a bivariate normal distribution with zero means and unit variances with correlation ρ_{01}. Then compute $T_1(1/2) = \omega_0 T_1(0) + \omega_1 T_1(1)$ and MAX3 denoted as MAX3$_j$. Hence, MAX3$_1$,. . ., MAX3$_m$ form an empirical null distribution of MAX3. We recommend using the method "asy" to compute the *P*-value of MAX3, especially for small *P*-values, which require a large number of replicates m. For example, in the illustration in Subheading 2.2, to obtain an accurate estimate of a *P*-value as small as 1e–8, we need at least 10 million replicates, which is computationally intensive. Using a smaller m, say, m = 100,000 the estimated *P*-value would be 0, which may be reported as "0" or as " < 2.2e–16" by R. In the following example, we used the "boot" procedure with 100,000 replicates. The output shows the *P*-value is less than 2.2×10^{-16}.

```
> MAX3(a, “boot”, 100000);
The MAX3 test using the asy method
data: a
statistic = 5.7587 p-value = <2.2e-16
```

One can also use an approximation of the tail probability of MAX3 to approximate the *P*-value of MAX3 (8), which has a closed form. This approximate *P*-value, however, is not reported by the R package Rassoc.

[3]Other Robust Tests for Binary Traits.

Nearly all robust methods have been developed for case–control studies. In addition to MAX3 (2), other robust tests are also developed for case–control association studies. A review of different robust tests for association studies can be found in ref. 9, 10. Although different robust tests have been developed, they have similar performance under the alternative hypothesis. In ref. 9, a function "casecontrol" in R is provided, which also outputs the three trend tests: Pearson's test, MAX3, and other robust tests. The *P*-values of the trend tests and Pearson's test are based on the asymptotic distributions, while the *P*-value for MAX3 is based on the bootstrap simulation. Discussion of applying single-marker analysis with robust tests in genome-wide association studies can be found in ref. 11.

$$F=\frac{SS_b/2}{SS_b/(n-3)}, \quad (3)$$

which under H_0, has an asymptotic F-distribution with (2,n – 3) df. We illustrate how to use these F-tests using some existing R functions later.

The allele-based analysis, valid under Hardy–Weinberg equilibrium (HWE), has similar performance to the genotype-based analysis under the additive model. Therefore, we focus on the genotype-based analysis, which does not require HWE. In Note 4, we present some power comparisons of different F-tests for quantitative traits.

2. Methods

2.1. Analysis of Case–Control Data

Denote the genotypes as (G_0, G_1, G_2) = (*AA*, *AB*, *BB*). If we happen to know the risk allele, it is always denoted as *B*. The choice of test statistic depends on which of the following four situations holds: (1) the genetic model and the risk allele are known, (2) the genetic model is known but not the risk allele, (3) the risk allele is known but not the genetic model, and (4) neither the genetic model nor the risk allele is known. Common genetic models include recessive, additive, and dominant. It is important that one does not determine the genetic model and/or the risk allele from the same data that will be used in the subsequent association analysis. Of course, the genetic model and/or the risk allele may be known based on scientific knowledge or information from previous data. In our view, (3) and (4) are the most common situations in practice.

2.1.1. Which Test to Use?—Which test to choose depends on each of the four situations outlined above (see Note 1). Let $\chi_2^2(1-\alpha)$ be the upper 100(1 – α)th percentile of χ_2^2 and $z(1-\alpha)$ be the upper 100(1 – α)th percentile of $N(0,1)$.

1. The genetic model and the risk allele are known. $T_1(x)$ is optimal and should be used. Since the risk allele is known, a one-sided H_1 is used and $z(0.95) = 1.645$. For the recessive (additive, or dominant) model, reject H_0 if $T_1(0) > z(0.95)$ ($T_1(1/2) > z(0.95)$, or $T_1(1) > z(0.95)$). In each case, the P-value equals the probability of $Z > T_1(x)$, where $Z \sim N(0,1)$.
2. The genetic model is known but not the risk allele. When the model is additive, use $T_1(1/2)$ and reject H_0 if $|T_1(1/2)| > z(0.975) = 1.96$. The P-value equals two times

---

[4]Comparison of *F*-Tests for Quantitative Traits.
We conducted a simulation to compare $F(x)$ (x = 0,1/2,1) and F by choosing the frequency of allele *B* (denoted as p), the sample size n, the heritability h, and the unit variance for the random error. Given a genetic model x and the values of p and h, we computed a and d. The empirical power is reported in Table 2. $F(x)$ is most powerful when x is correctly specified. However, when x is misspecified, $F(1/2)$ is most robust among the three $F(x)$ statistics. On the other hand, F is slightly less powerful than $F(1/2)$ under the additive or dominant models, but it protects against substantial power loss under the recessive model.

[1]Choosing among the trend tests, Pearson's test, and MAX3.
In practice, neither the genetic model nor the risk allele is known. The trend test $T_1(1/2)$ and Pearson's test T_2 are not robust when they are used alone. A robust test should protect against substantial loss of power when the model is misspecified (5, 6). To examine which test is most robust across the three genetic models, we conducted a simulation choosing the genotype relative risk (GRR), given by f_2/f_0, for a given x so that the optimal trend test for that x had about 80% power. The results are reported in Table 1. The power of the optimal test given a genetic model is in bold. The minimum power of each test across the three genetic models is presented. In the table, $p = \Pr(B)$. The test with higher minimum power for any of the three possible underlying genetic models is the most robust test. The results show that the power of $T_1(1/2)$ ranges from 30 to 80% for p = 0.1, 50 to 80% for p = 0.3, and 60 to 80% for p = 0.45. However, MAX3 is most robust as the power of MAX3 always exceeds 70% regardless of the underlying genetic model or the allele frequency p. The minimum power of T_2 across the three genetic models exceeds 70%, although it has slightly lower power than MAX3 in the simulation studies. More extensive simulations and results can be found in ref. 7.

the probability of $Z > |T_1(1/2)|$. When the model is recessive or dominant, use MAX3 (see (3) next).

3. The genetic model is unknown (regardless of the risk allele). Use MAX3. Three approaches are available to calculate the *P*-value of MAX3 using the R package Rassoc. But we recommend using the one based on the asymptotic null distribution of MAX3.

4. The same as (3). Thus, we only discuss (3) in the following.

Note that we do not recommend T_2, because it is always less powerful than MAX3 (see Note 1). If T_2 is used, reject H_0 if $T_2 > \chi^2_2(0.95) = 5.9915$. The *P*-value equals the probability of $T > T_2$, where $T \sim \chi^2_2$.

2.1.2. Examples Using R—The R package Rassoc can be loaded from

```
> library(Rassoc);
```

There are two functions **CATT(data,*x*)** and **MAX3(data, method,*m*)** in the package for computing the trend tests and MAX3 and their *P*-values. Pearson's test and its *P*-value can be obtained using an existing R function. In both functions, the "data" comprises a 2 × 3 contingency table, i.e., genotype counts (r_0, r_1, r_2) for cases and (s_0, s_1, s_2) for controls. In the first function, "*x*" is 0, 0.5, or 1 for the recessive, additive, or dominant models, respectively. In the second function, the "method" refers to the procedure to calculate the *P*-value of MAX3. Three methods are available: "boot" for the bootstrap procedure, "bvn" for the bivariate normal procedure, or "asy" for the asymptotic procedure.

The first two procedures are simulation-based and the last one is based on the asymptotic distribution of MAX3 (see Note 2). The "*m*" in the second function refers to the number of replicates when "boot" or "bvn" is used. When "asy" is used, "*m*" can be any positive integer.

For illustration, we use a SNP (rs420259) reported by the WTCCC (4), which was the only SNP showing strong association with bipolar disorder in a genome-wide association study (GWAS) with 500,000 SNPs (the actual number of SNPs tested after quality control steps is less than 500,000). The genome-wide significance level used by (4) was 5×10^{-7} for strong association. The genotype counts are (r_0, r_1, r_2) = (83, 755, 1020) and (s_0, s_1, s_2) = (260, 1134, 1537). The data can be entered as follows.

```
> = c(83,755,1020);
> s = c(260,1134,1537);
> a = matrix(rbind(r,s), nrow = 2, byrow = FALSE);
```

To check that the data are correctly entered, just type **a**.

```
> a;
83 755 1020
260 1134 1537
```

We may not have sufficient scientific knowledge to claim a priori which allele (*A* or *B*) is the risk one and what the true genetic model is. For illustration purpose, let us say we know a priori the true model is dominant and that *B* is the risk allele. The analysis is carried out based on the three situations outlined before.

1. If we know the genetic model is dominant and the risk allele is *B*, apply $T_1(1)$ using the R function **CATT** as follows.

```
> CATT(a,1);
The Cochran-Armitage trend test
data: a
statistic = 5.7587 p-valve = 8.478e-09
```

The output shows that $|T_1(1)| = 5.7587$ and its *P*-value is 8.478×10^{-9}. This is a two-sided test. We use a one-sided test because the risk allele is known. Thus, the actual *P*-value is half of the reported one, that is, 4.239×10^{-9}, which is less than the significance level 5×10^{-7}. Hence we reject H_0.

2 If we know the genetic model is dominant but not the risk allele, use MAX3 not $T_1(1)$. Suppose we happen to enter the data as **b** and apply $T_1(1)$ as before.

```
> r = c(1020,755,82);
> s = c(1537,1134,260);
> b = matrix(rbind(r,s), nrow = 2, byrow = FALSE);
> CATT(b,1);
The Cochran-Armitage trend test
data: b
statistic = 1.6618 p-valve = 0.09656
```

The two-sided *P*-value is 0.09656, which is not significant. This example shows that knowing the risk allele is necessary for using $T_1(0)$ or $T_1(1)$. The use of MAX3 is illustrated in (3) later. We first show how to use the following R function to obtain Pearson's test $T_2 = 33.165$ and its *P*-value 6.285×10^{-8}. This *P*-value is also significant but larger than that of $T_1(1)$ in case (1), because $T_1(1)$ is optimal for the dominant model. If we apply T_2 to the dataset **b**, we would obtain the same results.

```
> chisq.test(a);
Pearson's Chi-squared test
data: a
X-squared = 33.165, df = 2, p-value = 6.285e-08
```

3 If we do not know the genetic model, we apply MAX3 and calculate its *P*-value using the “asy” procedure. The reported statistic is MAX3 = 5.7587 with *P*-value 2.347×10^{-8}. Thus, we reject H_0. Note that this *P*-value is smaller than that of T_2 but larger than that of $T_1(1)$ in case (1).

```
> MAX3(a, “asy”, 1);
The MAX3 test using the asy method
```

```
data: a
statistic = 5.7587 p-value = 2.347e-08
```

If $x = 0.5$ is used in $T_1(x)$ regardless of the true genetic model and the risk allele, the following results show that the P-value of $T_1(1/2)$ is not significant.

```
> CATT(a,0.5);
The Cochran-Armitage trend test
data: a
statistic = 3.6966 p-value = 0.0002185
```

2.2. Quantitative Trait

2.2.1. Which Test to Use?—For a continuous trait, the four situations outlined before also apply. Let $F_{u,v}(1-\alpha)$ be the upper $100(1-\alpha)$th percentile of an F-distribution with (u,v) df.

1. When the genetic model and the risk allele are known, the statistic $F(x)$ given in Eq. 2 is used. Since the risk allele is known, a one-sided H_1 is used. For the recessive (additive, dominant) model, reject H_0 if $F(0) > F_{1,n-2}(0.95)$ ($F(1/2) > F_{1,n-2}(0.95)$, $F(1) > F_{1,n-2}(0.95)$), where $F(0)$ ($F(1/2)$, $F(1)$) is the observed statistic. In each case, the P-value equals half of the probability of $f_{1,n-2} > F(x)$, where $f_{1,n-2}$ follows an F-distribution with $(1,n-2)$ df.
2. The genetic model is known but not the risk allele. When the genetic model is additive, $F(x)$ given in Eq. 2 is used (with $x = 1/2$). Reject H_0 if $F(1/2) > F_{1,n-2}(0.95)$, where $F(1/2)$ is the observed statistic. The P-value equals the probability of $f_{1,n-2} > F(1/2)$. When the genetic model is not additive (either recessive or dominant), F given in Eq. 3 is used. Reject H_0 if $F > F_{2,n-3}(0.95)$, where F is also the observed statistic.

The P-value equals the probability of $f_{2,n-3} > F$, where $f_{2,n-3}$ follows an F-distribution with $(2,n-3)$ df.

3 When the genetic model is unknown, F given in Eq. 3 is used. The rejection rule and P-value are similar to those in case (2) when F is used.

4 The same as (3). Thus, we focus on (3) only.

2.2.2. Examples Using R—For illustration, we simulated a dataset called “QTLex.txt,” which contains (Y, G) for $n = 100$ individuals. In the simulation, the true model was dominant and the risk allele was B with population frequency 0.3. HWE was assumed in the population. The heritability was set to 0.1. The trait Υ was simulated from a normal distribution using the model given in Subheading 1 where $\mu = 0$, $E(\in) = 0$, and $\mathrm{Var}(\in) = 1$. The genotype G is AA, AB, or BB. The data can be read as follows.

```
> c = read.table(“QTLex.txt”,header=T)
```

If we know the true genetic model (dominant) and the risk allele a priori, we use $F(x)$ given in Eq. 2 with $x = 1$ as follows.

```
> objF1=aov(Y ~ (as.integer(G)==1),data=c);
```

```
> summary(objF1);
Df Sum Sq Mean Sq F value Pr($>$F)
as.integer(G) == 1 1 30.230 30.2295 31.317 1.993e-07 ***
Residuals Df 98 94.599 0.9653
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

The function “as.integer(G)” assigns 1 for *AA*, 2 for *AB*, and 3 for *BB*. Thus, “objF1 = aov(Y ~ (as.integer(G)==1),data = c)” conducts the ANOVA comparing the mean trait values between the two genotype groups: *AA* and *AB* + *BB*. In the first line of the output, “*F* value” is the statistic *F*(1) and “Pr(>F)” is the *P*-value. These values are reported in the second line. In this example, *F*(1) = 31.317 and the *P*-value is 1.993×10^{-7}. The strength of association is indicated by “***” near the *P*-value, and the interpretation of this significance code is given in the last line of the output. Since the risk allele is known, a one-sided test should be used. Thus, the actual *P*-value is half of that reported one, i.e., 9.965×10^{-8}. This *P*-value is very significant compared to the 0.05 significance level.

If we did not know the genetic model, the statistic *F* given in Eq. 3 should be used. See the following output. In this case, we would not assign scores (1, 2, 3) to the three genotypes. The output given below shows *F* = 15.575 with *P*-value 1.361×10^{-6}, which is also significant, but larger than the *P*-value obtained from *F*(1), which is the most powerful test when the true model is dominant.

```
> objF=aov(Y ~ G,data=c);
> summary(objF);
Df Sum Sq Mean Sq F value Pr(>F)
G 2 30.343 15.1715 15.575 1.361e-06 ***
Residuals 97 94.485 0.9741
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

For illustration, we also calculate *F*(1/2) and *F*(0) and their *P*-values. *F*(1/2) can be obtained by

```
> objF05=aov(Y~ (as.integer(G),data=c);
> summary(objF05);
```

In this case, “as.integer(G)” is equivalent to using scores 1, 2, 3 for the three genotypes (under the additive model). The reported *P*-value is 1.011×10^{-6}, which is the *P*-value if we do not know the risk allele. If we know the risk allele, the *P*-value is $1.011 \times 10^{-6}/2 = 5.055 \times 10^{-7}$. Both one-sided and two-sided *P*-values are significant. Interestingly, if we apply *F*(1/2) even when the risk allele is unknown, the two-sided *P*-value (1.011×10^{-6}) is smaller than that of *F*. The test *F*(0), which is optimal for arecessive model, is obtained as follows.

```
> objF0=aov(Y ~ (as.integer(G)==3),data=c);
> summary(objF0);
```

In this case, the ANOVA is applied with two genotype groups: *AA* + *AB* and *BB*. The reported *P*-value is 0.1398. If we know the risk allele, the *P*-value is 0.1398/2 = 0.0699, which is not significant at the 0.05 level. This illustrates loss of power can occur if the test used is not appropriate for the underlying genetic model.

References

1. Sasieni PD. From Genotypes to Genes: Doubling the Sample Size. Biometrics. 1997; 53:1253–1261. [PubMed: 9423247]
2. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case–control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002; 53:146–152. (Erratum (2009) 68: 220). [PubMed: 12145550]
3. Zang Y, Fung WK, Zheng G. Simple algorithms to calculate asymptotic null distributions of robust tests in case–control genetic association studies in R. J Stat Softw. 2010; 33(8):1–24. [PubMed: 20808728]
4. The Wellcome Trust Case Control Consortium (WTCCC). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
5. Gastwirth JL. On Robust Procedures. J Am Stat Assoc. 1966; 61:929–948.
6. Gastwirth JL. The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis. J Am Stat Assoc. 1985; 80:380–384.
7. Zheng, G.; Freidlin, B.; Gastwirth, JL. Comparison of robust tests for genetic association using case–control studies.. In: Rojo, J., editor. Optimality: The Second Eric L. Lehmann Symposium, IMS Lecture Notes-Monograph Series. Institute of Mathematical Statistics; Beachwood, Ohio: 2006.
8. Li QZ, Zheng G, Li Z, Yu K. Efficient approximation of *P*-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet. 2008; 72:397–406. [PubMed: 18318785]
9. Joo J, Kwak M, Chen Z, Zheng G. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. Stat Med. 2010; 29:158–180. [PubMed: 19918942]
10. Kuo CL, Feingold E. Wha's the Best Statistic for a Simple Test of Genetic Association in a Case–control Study? Genet Epidemiol. 2010; 34:246–253. [PubMed: 20025064]
11. Zheng G, Joo J, Tian X, Wu CO, Lin J-P, Stylianou M, Waclawiw MA, Geller NL. Robust genome-wide scans with genetic model selection using case–control design. Stat Its Interface. 2009; 2:145–151.

Table 1

Empirical power (%) and robustness of different tests for the analysis of case–control data

P	*x*	GRR	$T_1(0)$	$T_1(1/2)$	$T_1(1)$	T_2	MAX3
0.10	0.0	3.15	**80.40**	30.30	11.56	70.58	72.42
	0.5	1.88	21.23	**80.30**	79.20	73.09	76.89
	1.0	1.47	8.94	77.60	**79.85**	72.78	75.07
	Min		8.94	30.30	11.56	70.58	73.42
	Max of min						72.42
0.30	0.0	1.65	**80.37**	52.07	15.52	71.18	72.46
	0.5	1.60	44.78	**81.71**	76.59	73.01	76.96
	1.0	1.38	12.91	71.00	**80.73**	70.92	72.54
	Min		12.91	52.07	15.52	70.92	72.54
	Max of min						72.54
0.45	0.0	1.45	**79.69**	61.13	16.46	70.25	72.27
	0.5	1.57	56.52	**80.01**	67.91	71.52	76.07
	1.0	1.44	14.63	64.92	**80.81**	71.72	73.74
	Min		14.63	61.13	16.46	70.25	73.74
	Max of min						73.74

Methods Mol Biol. Author manuscript; available in PMC 2013 May 13.

Table 2

Empirical power (%) for the analysis of a quantitative trait using different test statistics given p, $h = 0.1$, and $n = 100$

Model (x)	P	$F(0)$	$F(1/2)$	$F(1)$	F
0.0	0.10	62.29	31.54	12.88	59.70
	0.30	87.65	56.16	15.60	80.49
	0.45	89.86	70.22	17.00	82.87
0.5	0.10	53.80	89.05	87.65	82.89
	0.30	57.15	90.57	83.96	83.50
	0.45	70.88	90.49	77.42	84.03
1.0	0.10	40.03	87.82	89.75	84.44
	0.30	15.58	84.14	90.15	83.39
	0.45	17.76	77.45	89.62	82.85

Methods Mol Biol. Author manuscript; available in PMC 2013 May 13.