Maximizing Association Statistics Over Genetic Models

Juan R. González,[1–3]* Josep L. Carrasco,[3] Frank Dudbridge,[4] Lluís Armengol,[2,5] Xavier Estivill,[2,5] and Victor Moreno[6]

[1] *Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain*
[2] *CIBER de Epidemiología y Salud Pública (CIBERESP), Spain*
[3] *Biostatistic Unit, Department of Public Health, University of Barcelona, Spain*
[4] *MRC Biostatistics Unit, Cambridge, United Kingdom*
[5] *Genes and Disease Program, Center for Genomic Regulation, Barcelona, Spain*
[6] *IDIBELL, Catalan Institute of Oncology, Barcelona, Spain*

The assessment of the association between a candidate locus and a disease may require the assumption of an inheritance model. Most researchers select the additive model and test the association with the Cochran-Armitage trend test. This test assumes a dose-response effect with regard to the number of copies of the variant allele. However, if there is reason to expect dominance or recessiveness in the effect of the variant allele, the heterozygous genotype may be grouped with one of the two homozygous, depending on the inheritance model, and a simple test on the 2×2 table can be used to assess independence. When the underlying genetic model is unknown, association may be assessed using the max-statistic, which selects the largest test statistic from the dominant, recessive and additive models. The statistical significance of the max-statistic has been previously addressed using permutation or Monte Carlo simulation approaches. We aimed to provide simpler alternatives to the max-test to make it feasible in large-scale association studies. Our simulations show that this procedure has an effective number of tests of 2.2, which can be used to correct the significance level or *P*-values. We also derive the asymptotic distribution of max-statistic, which leads to a simple way to calculate the significance level and allows the derivation of a formula for power calculations in the design of studies that plan to use the max-statistic. A simulation study shows that the use of the max-statistic is a powerful approach that provides safeguard against model uncertainty. *Genet. Epidemiol.* 32:246–254, 2008. © 2008 Wiley-Liss, Inc.

Key words: genetic models; max-statistic; mode of inheritance; power

The Supplemental Materials described in this article can be found at www.interscience.wiley.com/jpages/0741-0395/suppmat
Contract grant sponsor: CIBER de Epidemiología y Salud Pública; Contract grant number: CIBERESP-MET-002; Contract grant sponsors: Instituto de Salud Carlos III; The Spanish Secretaria de Estado de Universidades e Investigacion; Contract grant number: PR2006-0474; Contract grant sponsor: EU Sixth Framework Programme; Contract grant number: LSHM-CT-2004-503485.

*Correspondence to: Juan R. González, Center for Research in Environmental Epidemiology (CREAL) (Room 188), Barcelona Biomedical Research Park (PRBB), Plaza Charles Darwin s/n, Barcelona 08003, Spain. E-mail: jrgonzalez@imim.es

Received 19 July 2007; Revised 24 October 2007; Accepted 28 November 2007
Published online 28 January 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20299

INTRODUCTION

The recent popularization of high-throughput genotyping platforms has allowed researchers to undertake genome-wide scans for disease associated markers. The association between a candidate locus and a disease is often assessed using the genotype-based Cochran-Armitage (CA) trend test [Armitage, 1955]. This test assumes a dose-response effect with regard to the number of copies of the variant allele and is also called an additive model. However, if there is reason to expect dominance or recessiveness in the effect of the variant allele, the heterozygous genotype may be grouped with one of the two homozygous, depending on the risk model, and a simple test based on the 2×2 table can be used to assess independence.

Since assuming a model different from the real one leads to loss of power [Slager and Schaid, 2001; Freidlin et al., 2002; Schaid et al., 2005], some authors have proposed assessing the association using the largest test statistic from dominant, recessive and additive models [Milne et al., 2006; Pooley et al., 2006; Cargill et al., 2007]. This statistic is known as the max-statistic [Freidlin et al., 2002]. A naive approach is to consider the smaller *P*-value of χ^2_D, χ^2_R and χ^2_A as the significance of association between the single nucleotide polymorphism (SNP) and the disease. However, this approach does not maintain the overall type-I error rate as it does not account for multiple testing, as we will illustrate in the Results section.

To accurately overcome the multiple testing problem, the distribution of max-statistic should be used to calculate significance. Sladek et al. [2007] and Freidlin et al. [2002] pointed out that this distribution is unknown and used a permutation approach. However, this procedure may be extremely expensive in computational terms if applied in the context of genome-wide scans where thousands of SNPs are analyzed. As an example, Sladek et al. [2007] needed to calculate around 11,800 million tests (392,935 markers, three models and 10,000 permutations) to identify novel risk loci for type-2 diabetes in a genome-wide study. The computing time may increase dramatically when more permutations are needed to better estimate the distribution of max-statistic for those SNPs that are strongly associated with the disease, as the *P*-values in that case are in the tail of the distribution. This is what Sladek et al. [2007] did in a second stage where 10,000,000 permutations were carried out to better estimate the significance. All this computational burden could be reduced to just one evaluation of the three models if the distribution for the max-statistic were known. Also, this knowledge would be useful in the design phase of the studies to correctly calculate the sample size for a given power taking into account the inflated error rate when the max-statistic is used. Therefore, the main aim of this paper is to derive the asymptotic form for max-statistic as well as to study its finite properties using both simulated and real data sets.

METHODS

NOTATION

The *G*-test is a likelihood ratio test that can be used to evaluate the statistical significance for a given SNP [Cordell and Clayton, 2002]. The *G*-test compares the likelihood scores of the model including the SNP to the null model. For the general contingency table, we can write the statistic *G* as

$$G = 2\sum_i O_i \log(O_i/E_i),$$

where the sum is computed over all cells, i, and O denotes the observed number of cases and E the expected number of cases under the null hypothesis of no association. In a case-control study, for a general table $2 \times K$ (where K denotes the number of categories for independent variable) and using a similar notation given in Table I, the *G*-test may be expressed as follows:

$$\begin{aligned} G(\boldsymbol{p},\boldsymbol{q}) &= 2\sum_{i=1}^{K}\left\{ r_i \log\left(\frac{r_i N}{n_i R}\right) + s_i \log\left(\frac{s_i N}{n_i S}\right)\right\} \\ &= 2\sum_{i=1}^{K}\left\{ p_i R \log\left(\frac{2p_i}{p_i+q_i}\right) + q_i S \log\left(\frac{2q_i}{p_i+q_i}\right)\right\}, \end{aligned}$$

where $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{q} = (q_1, \ldots, q_K)$ are the sample proportions for cases and controls, respectively. In particular, for the case of a codominant model given in Table I, we have that $\boldsymbol{p} = (p_1, p_2, p_3)$ and $\boldsymbol{q} = (q_1, q_2, q_3)$. In this case, the *G* statistic follows a χ^2 distribution with 2 degrees of freedom.

TABLE I. Genotype distribution among cases and controls

	a/a	a/A	A/A	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Alternative simpler tests, based on one degree of freedom χ^2_1, may be defined assuming different effects for the risk allele based on inheritance models. For instance, the dominant model may be expressed as

$$\begin{aligned} \chi^2_D = G(\boldsymbol{p},\boldsymbol{q}) = 2\Bigg(& p_1 R \log\left[\frac{2p_1}{p_1+q_1}\right] + q_1 S \log\left[\frac{2q_1}{p_1+q_1}\right] \\ & + (p_2+p_3) R \log\left[\frac{2(p_2+p_3)}{p_2+p_3+q_2+q_3}\right] \\ & + (q_2+q_3) S \log\left[\frac{2(q_2+q_3)}{p_2+p_3+q_2+q_3}\right]\Bigg) \end{aligned} \tag{1}$$

and the recessive model corresponds to

$$\begin{aligned} \chi^2_R = G(\boldsymbol{p},\boldsymbol{q}) = 2\Bigg(& (p_1+p_2) R \log\left[\frac{2(p_1+p_2)}{p_1+p_2+q_1+q_2}\right] \\ & + (q_1+q_2) S \log\left[\frac{2(q_1+q_2)}{p_1+p_2+q_1+q_2}\right] \\ & + p_3 R \log\left[\frac{2p_3}{p_3+q_3}\right] + q_3 S \log\left[\frac{2q_3}{p_3+q_3}\right]\Bigg). \end{aligned} \tag{2}$$

Notice that in both cases the *G*-test corresponds to a simple 2×2 table.

The *G*-test for the additive model may not be expressed in a closed form but can be well approximated by the CA trend test, which can be expressed in terms of $\boldsymbol{p}$ and $\boldsymbol{q}$ as follows when the scores $(-1, 0, 1)$ are used:

$$\begin{aligned} \chi^2_A &= CA(\boldsymbol{p},\boldsymbol{q}) \\ &= \frac{(p_1-p_3-q_1+q_3)^2 RS(R+S)}{-[(p_1-p_3)R+(q_1-q_3)S]^2 + (R+S)[(p_1+p_3)R+(q_1+q_3)S]}, \end{aligned} \tag{3}$$

where $\boldsymbol{p} = (p_1, p_3)$ and $\boldsymbol{q} = (q_1, q_3)$.

ASYMPTOTIC DISTRIBUTION OF MAX-STATISTIC

The max-statistic may be expressed as

$$\chi^2_{\max} = \max\{\chi^2_D, \chi^2_R, \chi^2_A\}.$$

To obtain its asymptotic distribution, we will work with the statistic

$$T_{\max} = (-1)^s \max\{\sqrt{\chi_D^2}, \sqrt{\chi_R^2}, \sqrt{\chi_A^2}\},$$

where the positive square roots are taken, and $s = 1$ when $(r_2/r_0) < (s_2/s_0)$ (i.e., the sample odds ratio for A/A against a/a is less than one), 0 otherwise. Define $T_D = (-1)^s \sqrt{(\chi_D^2)}$, and T_R,T_A similarly.

As the square root is a monotonic function, the maximum χ^2 test will concur with the maximum square-root test. We notice that under independence $\Pr\{|T_{\max}| \leq m\} = \Pr\{|T_D| \leq m \cap |T_R| \leq m \cap |T_A| \leq m\} = \Pr\{\sqrt{\chi_1^2} \leq m\}^3$. Nonetheless, as we will later show in a simulation study, T_D, T_R and T_A are correlated because they are calculated from the same sample. Therefore, we need to compute $\Pr\{ | \mathrm{T}_{\max} | \leq \mathrm{m}\}$ using the joint distribution of T_D, T_R and T_A.

The delta method [Agresti, 2002] may be used to derive the joint distribution of the three test statistics, and hence, used to compute asymptotic P-values. In a typical case-control setting, we can consider that the data have been generated from two independent, multinomial distributions. One for cases with sample proportions $\boldsymbol{p} = (p_1,p_2,p_3)$ and other for controls with sample proportions $\boldsymbol{q} = (q_1,q_2,q_3)$. Let $\boldsymbol{\Xi_p}$ be the covariance matrix of $\boldsymbol{p}$ where $\boldsymbol{\Xi_p} = (\sigma_{pij})$ with

$$\sigma_{p_{jj}} = p_j(1 - p_j),$$

$$\sigma_{p_{jk}} = -p_j p_k \quad \text{for } j \neq k$$

and $\boldsymbol{\Xi_q}$ be the covariance matrix of $\boldsymbol{q}$ with

$$\sigma_{q_{jj}} = q_j(1 - q_j)$$

$$\sigma_{q_{jk}} = -q_j q_k \quad \text{for } j \neq k.$$

The $\boldsymbol{\Xi_p}$ and $\boldsymbol{\Xi_q}$ matrices have the forms

$$\boldsymbol{\Xi_p} = [\mathbf{diag}(\boldsymbol{p}) - \boldsymbol{pp}']$$

and

$$\boldsymbol{\Xi_q} = [\mathbf{diag}(\boldsymbol{q}) - \boldsymbol{qq}'],$$

respectively. Therefore, the full covariance matrix, $\boldsymbol{\Xi}$, of the observed proportions in the contingency table will be

$$\boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{\Xi_p} & 0 \\ 0 & \boldsymbol{\Xi_q} \end{pmatrix}.$$

If $\boldsymbol{\pi}$ is the vector of sample proportions, $\boldsymbol{\pi} = (\boldsymbol{p},\boldsymbol{q}) = (p_1,p_2,p_3,q_1,q_2,q_3)$, the joint distribution of $\boldsymbol{\pi}$ is asymptotically multivariate normal [Agresti, 2002]. The delta method generalizes further to a vector of functions of an asymptotically normal random vector. Therefore, as $T_{\max}$ is a vector of functions of $\boldsymbol{\pi}$, which is asymptotically normal, we obtain that $T_{\max}$ follows a trivariate normal distribution with asymptotic variance given by

$$\phi(H_{\mathrm{i}})'\boldsymbol{\Xi}\phi(H_{\mathrm{i}}) \tag{4}$$

and covariances for each pair as

$$\phi(H_{\mathrm{i}})'\boldsymbol{\Xi}\phi(H_{\mathrm{j}}), \tag{5}$$

where

$$\phi(H_{\mathrm{i}}) = \frac{\partial H_{\mathrm{i}}}{\partial \pi_c}, \quad c = 1, \ldots, 2K.$$

π_c indicates the different components of the vector $\boldsymbol{\pi}$, $H_1 = T_D$, $H_2 = T_{\mathrm{R}}$ and $H_3 = T_{\mathrm{A}}$ and K is the number of columns considered in the $2 \times K$ table (in our case $K = 3$). Expressions for the partial derivatives for each of the three modes of inheritance are given in Appendix A.

Noting further that each of T_D, T_R, T_A has mean zero under no association, we can compute $\Pr\{|T_{\max}| \leq m\}$ using the trivariate normal distribution, N_3, as follows:

$$\int_{-m}^{m}\int_{-m}^{m}\int_{-m}^{m} N_3(\underline{z}; \underline{0}, \boldsymbol{\Sigma})\, d\underline{z}, \tag{6}$$

where $m = \max\{\sqrt{\chi_D^2}, \sqrt{\chi_R^2}, \sqrt{\chi_A^2}\}$ and $\underline{z} = (z_1, z_2, z_3)$ are multivariate normal variables with the following covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{T_D T_R} & \rho_{T_D T_A} \\ \rho_{T_R T_D} & 1 & \rho_{T_R T_A} \\ \rho_{T_A T_D} & \rho_{T_A T_R} & 1 \end{pmatrix}, \tag{7}$$

where the correlations can be computed using (4) and (5) as follows:

$$\rho_{T_D T_R} = \frac{\phi(T_D)'\boldsymbol{\Xi}\phi(T_R)}{[\phi(T_D)'\boldsymbol{\Xi}\phi(T_D)]^{1/2}[\phi(T_R)'\boldsymbol{\Xi}\phi(T_R)]^{1/2}}$$

$$\rho_{T_D T_A} = \frac{\phi(T_D)'\boldsymbol{\Xi}\phi(T_A)}{[\phi(T_D)'\boldsymbol{\Xi}\phi(T_D)]^{1/2}[\phi(T_A)'\boldsymbol{\Xi}\phi(T_A)]^{1/2}}$$

and

$$\rho_{T_R T_A} = \frac{\phi(T_R)'\boldsymbol{\Xi}\phi(T_A)}{[\phi(T_R)'\boldsymbol{\Xi}\phi(T_R)]^{1/2}[\phi(T_A)'\boldsymbol{\Xi}\phi(T_A)]^{1/2}}$$

Equation (6) may be computed using numerical integration as described in Genz [1992]. Notice that delta method may fail near singularity points. This may happen with sparse data or with rare mutations. In this case, the theorem given by Puig [1998], which is an extension of the one published in Serfling [1980], may be used to avoid problems with singularity.

These expressions may appear complicated, but they can be easily evaluated for a given 2×3 table. We have implemented such calculations in R code, which is available in our analysis R package SNPassoc [Gonzalez et al., 2007] (freely available at http://davinci.crg.es/estivill_lab/snpassoc). For practical purposes, we have given in Table IV some quantiles of the asymptotic distribution for max-statistic under the assumption of Hardy-Weinberg equilibrium (HWE) and for different allelic frequencies.

POWER OF MAX-STATISTIC

Power calculations are important to design good genetic studies. Herein, we derive the formula for power for studies that plan to use the max-statistic for the analysis. The formula assumes independent ascertainment of cases and controls and does not make any assumption among the distribution of genotypes. In other words, the number of cases and controls is fixed and HWE for genotype counts may not hold. These assumptions are the same as Slager and Schaid [2001] considered to give a formula for the power when Armitage's trend test is used.

Power calculations for tests based on the χ^2 distribution are based on a non-centrality parameter, λ [Guenther, 1977]. For the G-test, the non-centrality parameter was given by Agresti [2002]. In our notation it takes the form

$$\lambda = 2\left(R\sum_{i=1}^{3}\left\{ p_i \log\frac{2p_i}{p_i+q_i}\right\} + S\sum_{i=1}^{3}\left\{ q_i \log\frac{2q_i}{p_i+q_i}\right\}\right),$$

where p_i, $i=1,2,3$, depend on the allelic frequency and they are assumed to be in HWE and q_i, $i=1,2,3$, depend on both the allelic frequency and the expected difference among cases and controls (measured as an odds ratio, OR). Thus, the non-centrality parameter depends on the allelic frequency, OR and the sample sizes of cases and controls. This parameter may be obtained for the simplified models χ^2_D and χ^2_R changing the probabilities p_i and q_i as in (1) and (2), respectively. Regarding the CA trend test, e.g., χ^2_A based on Chapmand and Nam's [1968] work, it can be shown that the non-centrality parameter is given by

$$\lambda_A = RS\frac{[\sum_{i=1}^{3} x_i(p_i-q_i)]^2}{\sum_{i=1}^{3} x_i^2(Rp_i+Sq_i) - [\sum_{i=1}^{3} x_i(Rp_i-Sq_i)]^2/N}, \tag{8}$$

where x_i are the scores that in our case were considered as $(-1,0,1)$.

Under the alternative hypothesis of association, each χ^2 is distributed as a $\chi^2_{1,\lambda}$, which has the same distribution as a squared $N(\sqrt{\lambda},1)$. We assume that the association is parameterized by a single OR under one of the above models, in which case T_D, T_R, T_A are positively correlated. Therefore, the $p\{|T_{\max}|\leq m\}$ under the alternative hypothesis is given by

$$\int_{-m}^{m}\int_{m}^{m}\int_{-m}^{m} N_3(\underline{z};\underline{\mu},\mathbf{\Sigma})\mathrm{d}\underline{z}, \tag{9}$$

where $m=\max\{\sqrt{\chi^2_D},\sqrt{\chi^2_R},\sqrt{\chi^2_A}\}$, $\underline{z}=(z_1,z_2,z_3)$, $\underline{\mu}=(\sqrt{\lambda_D},\sqrt{\lambda_R},\sqrt{\lambda_A})$, λ_D, λ_R and λ_A are the non-centrality parameters for T_D, T_R and T_A, respectively, and $\mathbf{\Sigma}$ is the correlation matrix given in Equation (7). It then follows that the asymptotic power is calculated as

$$p\{T_{\max_{Ha}} > T_{\max_{H0}}(\alpha)\},$$

where $T_{\max_{Ha}}$ makes reference to the distribution of $T_{\max}$ under the alternative hypothesis (Equation (9)) and $T_{\max_{H0}}(\alpha)$ is the critical value at α-level of $T_{\max}$ under the null hypothesis (Equation (6)).

RESULTS

SIMULATION STUDY

As we have mentioned in the Methods section, some authors consider the significance level of the max-statistic as the P-value corresponding to the smaller P-value of χ^2_D, χ^2_R and χ^2_A. However, this approach does not maintain the overall type-I error rate as it does not account for multiple testing. To illustrate this problem, we simulated case-control data from the null hypothesis (e.g. $p_i=q_i$, $i=1,2,3$) assuming that controls were in HWE, which is usually expected for genetic data, though simulations assuming that HWE did not hold were also carried out leading to the same conclusions (data not shown). The sample sizes for the simulated data were selected from the range {100–300} and the frequency for the minor allele in the range {0.1–0.5}. For each scenario, the type-1 error rate was estimated after performing 20,000 simulations.

Effective number of tests. Table II shows the proportion of tests that rejected the null hypothesis at the 5% level after considering the minimum P-value of χ^2_D, χ^2_R and χ^2_A as a criterion. As expected, the type-1 error rate is clearly increased because this procedure is based on performing three tests at a time. If the three tests were independent, one would expect these simulated type-1 error rates to be around 15%. However, due to the correlation among the tests, we observe that the values are around 10.5%. This indicates that the effective number of

TABLE II. Achieved levels of each of the 5%-asymptotic level tests based on 20,000 replications taking the most significant of χ^2_A, χ^2_D and χ^2_R test as a signification value

n	0.5	0.3	0.2	0.1
100	10.79	11.17	12.39	12.16
200	10.59	11.22	11.42	13.41
300	10.68	10.12	10.46	13.56
400	10.51	10.31	10.56	11.98
500	10.42	10.82	10.59	11.69
1,000	10.69	10.33	10.62	11.24
2,000	10.37	10.68	10.51	10.64
2,500	10.55	10.44	10.46	10.91
3,000	10.21	10.72	10.12	10.68

TABLE III. Some quantiles for max-statistic assuming HWE

Frequency of allele *A*	Max-statistic at nominal significance level 0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
0.5	5.208	7.964	12.118	16.446	20.840	25.264	29.717
0.3	5.281	8.091	12.156	16.449	20.839	25.262	29.710
0.2	5.368	8.150	12.182	16.452	20.790	25.219	29.619
0.1	5.430	8.259	12.275	16.503	20.789	25.087	29.448
0.05	5.461	8.282	12.366	16.516	20.624	24.601	28.893

HWE, Hardy-Weinberg equilibrium.

TABLE IV. Achieved levels of each of the 5%-asymptotic level tests based on 20,000 replications using the asymptotic distribution of the max-statistic to get the significance level

n	0.5	0.3	0.2	0.1
100	4.79	4.82	4.70	5.62
200	4.88	4.86	4.72	5.78
300	4.79	4.79	5.19	5.21
400	4.82	5.12	5.26	4.74
500	5.11	5.07	5.12	4.80
1,000	5.04	4.94	5.07	4.82
2,000	5.02	5.06	4.89	5.21
2,500	4.96	5.08	5.06	5.08
3,000	4.98	4.97	4.95	5.06

tests is less than 3, but clearly greater than 1, and that a multiple testing correction should be done. The effective number of tests, k, can be estimated to be 2.2 after applying Şidàk's [1967] formula that for type-1 error equal to 0.05 is estimated as $k = \log(1-0.105)/\log(1-0.05)$ (Table II). Therefore, we might initially consider using the corrected significance level $\alpha' = \alpha/2.2$ to assure an α nominal significance level when three genetic models are considered. In this case, as we are considering the nominal level as the 5%, the corrected α' would be 0.023 (= 0.05/2.2). This value corresponds to a max-statistic of 5.18, but as we can observe from Table III, this is a conservative value and ignores the differences depending on the allele frequency.

Coverage of max-statistic. To illustrate whether this derived distribution is reasonable, we have used the simulated case-control data previously described. As an example, Figure 1 shows the case of having the allelic frequency equal to 0.2. The results indicate a good agreement between the theoretical distribution and simulated data. To further illustrate the behavior of max-statistics we computed the type-I error rate for the same simulations previously described. The results, given in Table IV, show that the estimated type-1 error is as expected (e.g. close to 0.05). The observed variation in the empirical 5%-asymptotic level can be considered as random variability. Under normality (this can be assumed as we performed 20,000 runs) the empirical 5%-asymptotic should be in interval (4.69, 5.31%) with 95% confidence. The only result that does not fall into this interval is 5.62 and 5.78 that can be explained due to the limited number of cases and controls and low allelic frequency. Anyway, as we are performing 36 estimations, we would expect to have about two outside this interval.

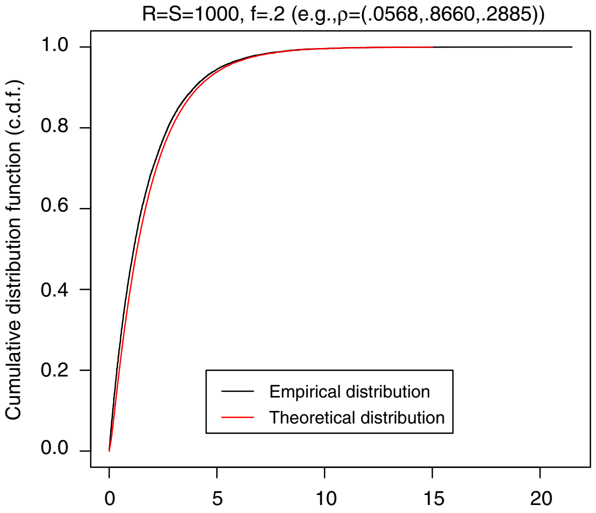

Fig. 1. Theoretical distribution of $\chi^2_{\max} = \max\{\chi^2_D, \chi^2_R, \chi^2_A\}$ (red line) and empirical distribution (black line) under the null hypothesis for a case-control study of sample size 1,000 in which $\rho_{\sqrt{\chi^2_D}\sqrt{\chi^2_R}} = 0.0568$, $\rho_{\sqrt{\chi^2_D}\sqrt{\chi^2_A}} = 0.8660$ and $\rho_{\sqrt{\chi^2_R}\sqrt{\chi^2_A}} = 0.2885$. These values correspond to simulated data in which the allelic frequency is set equal to 0.2.

Power of max-statistic. We have also carried out a simulation study to illustrate that the use of the max-statistic is a more sensitive method and provides a safeguard against model uncertainty. We have also compared the max-statistic with the three statistics based on likelihood ratio test (e.g. χ^2_D, χ^2_R and χ^2_A), the codominant model (χ^2_C) and the constrained-likelihood approach (Fig. 2). The statistic χ^2_C follows a χ^2 with 2 degrees of freedom, whereas the distribution of the constrained-likelihood statistic is given in Wang and Sheffield [2005]. As expected, χ^2_D performs best when the generating model is dominant, as do the statistic χ^2_R when the model is recessive and the statistic χ^2_A when the model is additive. That is, the maximum power is achieved when the true model is used. The statistic χ^2_D does not work well when the generating model is recessive, and the statistic χ^2_R is not recommended when the true model is either dominant or additive. Overall, max-statistic, constrained-likelihood and codominant tests seem to have a similar power. This indicates that they are

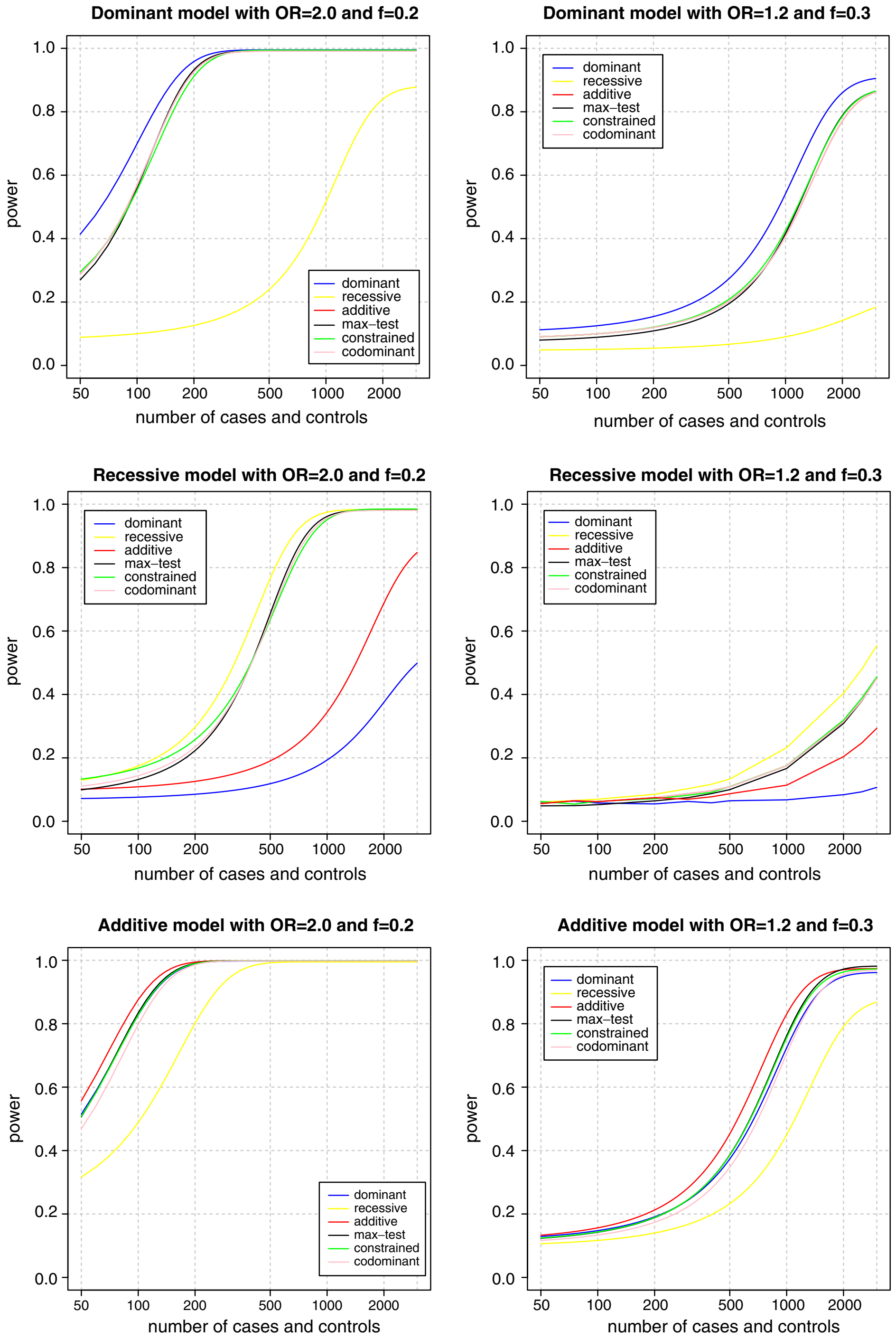

Fig. 2. Empirical power for max-statistic, constrained-likelihood, codominant, dominant, recessive and additive models. The results are given assuming different underlying genetic models for different odds ratio (2.0 and 1.2) and different allelic frequencies (0.2 and 0.3). The results are based on 20,000 replications.

good safeguards when the mode of inheritance is unknown. This fact is even more relevant when the underlying mode of inheritance is recessive, which has low power to be detected using an additive model. Nonetheless, when the underlying model is additive, the codominant test shows marginally worse power than the other two approaches.

Application to real studies. We have re-analyzed the genome-wide study carried out by Sladek et al. [2007] to identify risk loci for type-2 diabetes.

TABLE V. Genotype frequencies of the PGR-05 (rs660149) polymorphism in breast cancer from Pooley et al. [2006]

	G/G	G/C	C/C	Total
Cases	2,421	1,719	327	4,467
Controls	2,353	1,855	334	4,542
Total	4,774	3,574	661	9,009

The P-values computed using the proposed asymptotic distribution, compared with those obtained by the authors using a permutation approach, are given in the supplementary Table S1. We observe an almost perfect agreement among the results, even when we are computing significance in the tail of the distribution. In addition, the little observed differences do not follow any pattern (e.g. due to major allele frequency (MAF)). They may be explained by the fact that the P-value computed using asymptotic distribution of max-statistic is not exact since the integral is approximated by numerical integration and this may lead to little differences between both methods.

Also, to further illustrate the performance of our method and to emphasize the need of using the max-statistic, we have re-analyzed some data from Pooley et al. [2006] who studied the association of the progesterone receptor gene with breast cancer. Assuming codominant and additive models, the authors concluded that only the variant rs1042638 was significantly associated with the disease. However, using the max-statistic we found that rs660149 variant is also associated with the disease (data are given in Table V). After fitting the dominant and recessive models, we obtain that the P-value for the dominant model is 0.0230 and the P-value for the recessive model is 0.9203. This means that the rs660149 polymorphism would also be associated with breast cancer assuming a dominant model and using the same arguments as the authors did (e.g. considering the smaller P-value of the tests). However, the correct P-value should be computed using the null distribution of max-statistic. The asymptotic distribution of max-statistic leads to a $P\text{-value} = 0.043$, indicating that this polymorphism was also associated with breast cancer. On the other hand, neither constrained-likelihood (5.17, $P\text{-value} = 0.05619$) nor codominant (5.59, $P\text{-value} = 0.061$) tests where statistically significant at 0.05 level. Regarding the power calculation using equations given in the power of max-statistic section, as we have 4,500 cases and controls, an allelic frequency of 0.26 and assuming an OR equal to 1.2, the power of using max-statistic is between 67%, if the unknown genetic model is recessive, and 99% if the data are generated under a dominant or an additive genetic model.

CONCLUSION

In analysis of case-control genetic association studies, the max-statistic is useful. Our simulations have shown that the statistical significance of this test based on naively considering the minimum P-value leads to an increased type-1 error. Thus, we emphasize that the significance should be addressed using the asymptotic distribution because both simulation and real data studies have shown an excellent performance of this statistic. In our simulations, the max-statistic performed as well or better than the codominant and constrained-likelihood approaches when the true model was purely additive, recessive or dominant. For intermediate models its performance might be slightly worse, but it is currently unclear to what extent such intermediate models are present in nature.

Our analytic distribution can be calculated much more rapidly than a permutation test. Note, however, that the formula needs the allele frequency, which is typically not known and is estimated from data. Our simulations included this estimation and showed that the right type-1 error rates were obtained in the long run. Calculation of tail probabilities for single data sets may be inaccurate when the allele frequency is poorly estimated, as would be the case in small samples or when the minor allele is very rare.

The formula given for sample size computation may be employed to design genetic studies for which max-statistic is planned to be used. All these procedures are implemented in the R package called SNPassoc that is designed to analyze whole genome association studies [Gonzalez et al., 2007]. This package also includes a permutation test needed to appropriately account for the multiple comparisons usually made in whole-genome association studies where thousands of markers are explored.

ACKNOWLEDGMENTS

We are grateful to Dr. Pere Puig for valuable conversations about the distribution of max-statistic and δ method and Dr. Edsel Peña for helpful feedback. Dr. Kai Wang is also acknowledged for sending us some R functions to compute constrained-likelihood approach. We thank the two reviewers for their comments, which have led to improvement of the paper. This work has been partially supported by Methods Program from CIBER de Epidemiología y Salud Pública (CIBERESP-MET-002) funded by Instituto de Salud Carlos III (J. R. G.), the Spanish Secretaria de Estado de Universidades e Investigacion, PR2006-0474 (V. M.), and EU Sixth Framework Programme, LSHM-CT-2004-503485 (F. D.).

REFERENCES

Agresti A. 2002. Categorical Data Analysis. New York: Wiley.

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Cargill M, Schrodi SJ, Chang M, Garcia VE, Brandon R, Callis KP, Matsunami N, Ardlie KG, Civello D, Catanese JJ, Leong DU, Panko JM, McAllister LB, Hansen CB, Papenfuss J, Prescott SM, White TJ, Leppert MF, Krueger GG, Begovich AB. 2007. A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. Am J Hum Genet 80:273–290.

Chapman DG, Nam JM. 1968. Asymptotic power of chi square tests for linear trends in proportions. Biometrics 24:315–327.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. Am J Hum Genet 70:124–141.

Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146–152.

Genz A. 1992. Numerical computation of multivariate normal probabilities. J Comput Graphical Stat 1:141–150.

Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23:654–655.

Guenther WC. 1977. Power and sample size for approximate chi-square tests. Am Stat 31:83–85.

Milne RL, Ribas G, Gonzalez-Neira A, Fagerholm R, Salas A, Gonzalez E, Dopazo J, Nevanlinna H, Robledo M, Benitez J. 2006. ERCC4 associated with breast cancer risk: a two-stage case-control study using high-throughput genotyping. Cancer Res 66:9420–9427.

Pooley KA, Healey CS, Smith PL, Pharoah PD, Thompson D, Tee L, West J, Jordan C, Easton DF, Ponder BA, Dunning AM. 2006. Association of the progesterone receptor gene with breast cancer risk: a single-nucleotide polymorphism tagging approach. Cancer Epidemiol Biomarkers Prev 15:675–682.

Puig P. 1998. A note on testing segregation between two groups of animals using entropy. Biometrical J 40:155–163.

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. 2005. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76:780–793.

Serfling RJ. 1980. Approximation Theorems of Mathematical Statistics. New York: Wiley.

Sidàk Z. 1967. Rectangular confidence region for the means of multivariate normal distributions. J Am Stat Assoc 62:626–633.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885.

Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum Hered 52:149–153.

Wang K, Sheffield VC. 2005. A constrained-likelihood approach to marker–trait association studies. Am J Hum Genet 77:768–780.

APPENDIX A: DERIVATIVES OF *G* STATISTICS

The derivatives needed to compute (4) are the following:

For the dominant model,

$$\phi(\sqrt{\chi_D^2}) = \frac{1}{2\sqrt{\chi_D^2}}\left(\frac{\partial\chi_D^2}{\partial p_1}, \frac{\partial\chi_D^2}{\partial p_2}, \frac{\partial\chi_D^2}{\partial p_6 3}, \frac{\partial\chi_D^2}{\partial p_4}, \frac{\partial\chi_D^2}{\partial p_5}, \frac{\partial\chi_D^2}{\partial p_6}\right),$$

where

$$\begin{aligned} \frac{\partial\chi_D^2}{\partial p_1} &= \frac{2q_1(R-S)}{p_1+q_1} + 2R\log\left(\frac{2p_1}{p_1+q_1}\right) \\ \frac{\partial\chi_D^2}{\partial p_2} = \frac{\partial\chi_D^2}{\partial p_3} &= \frac{2(q_2+q_3)(R-S)}{p_2+p_3+q_2+q_3} + 2R\log\left(\frac{2(p_2+p_3)}{p_2+p_3+q_2+q_3}\right) \\ \frac{\partial\chi_D^2}{\partial q_1} &= \frac{2p_1(S-R)}{p_1+q_1} + 2S\log\left(\frac{2q_1}{p_1+q_1}\right) \text{ and} \\ \frac{\partial\chi_D^2}{\partial q_2} = \frac{\partial\chi_D^2}{\partial q_3} &= \frac{2(p_2+p_3)(S-R)}{p_2+p_3+q_2+q_3} + 2S\log\left(\frac{2(q_2+q_3)}{p_2+p_3+q_2+q_3}\right). \end{aligned}$$

For the recessive model,

$$\phi(\sqrt{\chi_R^2}) = \frac{1}{2\sqrt{\chi_R^2}}\left(\frac{\partial\chi_R^2}{\partial p_1}, \frac{\partial\chi_R^2}{\partial p_2}, \frac{\partial\chi_R^2}{\partial p_3}, \frac{\partial\chi_R^2}{\partial p_4}, \frac{\partial\chi_R^2}{\partial p_5}, \frac{\partial\chi_R^2}{\partial p_6}\right)$$

where

$$\begin{aligned}\frac{\partial\chi_R^2}{\partial p_1}=\frac{\partial\chi_R^2}{\partial p_2}&=\frac{2(q_1+q_2)(R-S)}{p_1+p_2+q_1+q_2}+2R\log\left(\frac{2(p_1+p_2)}{p_1+p_2+q_1+q_2}\right)\\ \frac{\partial\chi_R^2}{\partial p_3}&=\frac{2q_3(R-S)}{p_3+q_3}+2R\log\left(\frac{2p_3}{p_3+q_3}\right)\\ \frac{\partial\chi_R^2}{\partial q_1}=\frac{\partial\chi_R^2}{\partial q_2}&=\frac{2(p_1+p_2)(S-R)}{p_1+p_2+q_1+q_2}+2S\log\left(\frac{2(q_1+q_2)}{p_1+p_2+q_1+q_2}\right)\quad\text{and}\\ \frac{\partial\chi_R^2}{\partial q_3}&=\frac{2p_3(S-R)}{p_3+q_3}+2S\log\left(\frac{2q_3}{p_3+q_3}\right).\end{aligned}$$

Finally, for the additive model,

$$\phi(\sqrt{\chi_A^2})=\frac{1}{2\sqrt{\chi_A^2}}\left(\frac{\partial\chi_A^2}{\partial p_1},\frac{\partial\chi_A^2}{\partial p_2},\frac{\partial\chi_A^2}{\partial p_3},\frac{\partial\chi_A^2}{\partial p_4},\frac{\partial\chi_A^2}{\partial p_5},\frac{\partial\chi_A^2}{\partial p_6}\right)$$

where

$$\begin{aligned}&\frac{\partial\chi_A^2}{\partial p_1}\\ &=\frac{-\big((p_1-p_3-q_1+q_3)RS(R+S)^2((-q_1+p_1(-1+2q_1-2q_3)+q_3+p_3(-3-2q_1+2q_3))R+2\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S)\big)}{p_1^2R^2+(-1+p_3)p_3R^2-(p_3+q_1+2p_3q_1+q_3-2p_3q_3)RS+\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S^2-p_1R(R+2p_3R+S-2q_1S+2q_3S)^2}\end{aligned}$$

$$\begin{aligned}&\frac{\partial\chi_A^2}{\partial p_3}\\ &=\frac{\big((p_1-p_3-q_1+q_3)RS(R+S)^2((q_1+p_1(-3+2q_1-2q_3)-q_3+p_3(-1-2q_1+2q_3))R+2\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S)\big)}{(p_1^2R^2+(-1+p_3)p_3R^2-(p_3+q_1+2p_3q_1+q_3-2p_3q_3)RS+\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S^2-p_1R(R+2p_3R+S-2q_1S+2q_3S))^2}\end{aligned}$$

$$\begin{aligned}&\frac{\partial\chi_A^2}{\partial q_1}\\ &=\frac{((p_1-p_3-q_1+q_3)RS(R+S)^2(2\big(p_1^2+(-1+p_3)p_3-p_1(1+2p_3)\big)R+(-p_1+p_3-q_1+2p_1q_1-2p_3q_1-3q_3-2p_1q_3+2p_3q_3)S))}{(p_1^2R^2+(-1+p_3)p_3R^2-(p_3+q_1+2p_3q_1+q_3-2p_3q_3)RS+\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S^2-p_1R(R+2p_3R+S-2q_1S+2q_3S))^2}\end{aligned}$$

$$\begin{aligned}&\frac{\partial\chi_A^2}{\partial q_3}\\ &=\frac{-((p_1-p_3-q_1+q_3)RS(R+S)^2(2p_1^2R+2(-1+p_3)p_3R-(p_3+3q_1+2p_3q_1+q_3-2p_3q_3)S+p_1(-2R-4p_3R+S+2q_1S-2q_3S)))}{(p_1^2R^2+(-1+p_3)p_3R^2-(p_3+q_1+2p_3q_1+q_3-2p_3q_3)RS+\big(q_1^2+(-1+q_3)q_3-q_1(1+2q_3)\big)S^2-p_1R(R+2p_3R+S-2q_1S+2q_3S))^2}\end{aligned}$$

and

$$\frac{\partial\chi_A^2}{\partial p_2}=\frac{\partial\chi_A^2}{\partial q_2}=0.$$