J Hum Genet (2007) 52:747–756
DOI 10.1007/s10038-007-0177-7

ORIGINAL ARTICLE

Entropy-based joint analysis for two-stage genome-wide association studies

Guolian Kang · Yijun Zuo

Received: 26 February 2007 / Accepted: 2 July 2007 / Published online: 9 August 2007

Abstract Genome-wide association studies (GWAS) are being conducted to identify common genetic variants that predispose to human diseases to unravel the genetic etiology of complex human diseases now. Because of genotyping cost constraints, it often follows a two-stage design, in which a large number of markers are identified in a proportion of the available samples in stage 1, and then the markers identified in stage 1 are examined in all the samples in stage 2. In this paper, we introduce a nonlinear entropy-based statistic for joint analysis for two-stage genome-wide association studies. Type I error rates and power of the entropy-based statistic for association tests are validated using simulation studies in single-locus test. The power of entropy-based joint analysis is investigated by simulations. And the results suggest that entropy-based joint analysis is always more powerful than linear joint analysis that uses a linear function of risk allele frequencies in cases and controls when detecting rare genetic variants; the powers of these two joint analyses are comparable when detecting common genetic variants. Furthermore, when the false discovery rate is controlled, entropy-based joint analysis is more powerful and needs fewer samples than linear joint analysis that uses a linear function of risk allele frequencies in cases and controls. So, we recommend we should use entropy-based strategy for two-stage genome-wide association studies to detect the rare and common genetic variants with moderate to large genetic effect underlying a complex disease.

G. Kang (✉) · Y. Zuo
Department of Statistics and Probability,
Michigan State University,
East Lansing, MI 48824, USA
e-mail: kangg@stt.msu.edu

Keywords Complex diseases · Entropy · False discovery rate · Genetic variants

Introduction

Genome-wide association studies that were first suggested a decade ago by Risch and Merikangas (1996) are being conducted to unravel the genetic etiology of complex human diseases (Klein et al. 2005; Thomas et al. 2005), enabled by rapidly decreasing genotyping costs, massively high throughout genotyping technologies, the large-scale SNP discovery and genotyping efforts of the SNP Consortium (Sachidanandam et al. 2001) and the international HapMap consortium (Hinds et al. 2005). Presently the two-stage design is a more efficient method for genome-wide association studies than one-stage design.

In one-stage design, all available samples are genotyped on all markers. In replication-based two-stage analysis, a dense set of SNP markers across the genome is genotyped and tested using a portion of the samples in stage 1, and, the most-promising markers are then genotyped and tested in the remaining samples in stage 2. Compared with these two designs, in a joint analysis for two-stage genome-wide association studies, the most promising markers identified in stage 1 are examined in all samples in stage 2. So, joint analysis is more efficient, and its power is nearly the same as that of the one-stage design, while substantially reducing genotyping costs (Satagopan et al. 2002, 2004; Satagopan and Elston 2003; Thomas et al. 2004; Skol et al. 2006). Now, the crucial and urgent task for two-stage genome-wide association analysis is to construct more powerful test statistics in order to make good use of data information and to develop more efficient methods.

The power of two-stage genome-wide association studies to identify variants underlying a complex disease depends on a number of factors, including how M markers are selected, how many samples (N) are selected, how samples (π_{samples}) are divided between stage 1 and stage 2, the proportion (π_{markers}) of markers tested in stage 2 and strategy used to test for association, inheritable disease models, effect sizes of risk allele and disease prevalence and so on (Skol et al. 2006). So, for different parameters controlled, the relationships among $\pi_{\text{samples}}, \pi_{\text{markers}}$ and N need to be determined in order to get the higher power and control the proper genome-wide type I error rate.

Skol et al. (2006) get that joint analysis using a linear function of risk allele frequencies in cases and controls is more powerful than replication-based analysis for two-stage genome-wide association studies. (This joint analysis in Skol et al. will be referred to as the "linear joint analysis" throughout the article.) This can be achieved easily because linear joint analysis examines the data from both stages 1 and 2 and not from only stage 2 in the second stage. However, the statistic in Skol et al. (2006) compares risk allele frequencies in cases and controls and uses a linear function of risk allele frequencies in cases and controls. A nonlinear function of risk allele frequencies in cases and controls is Shannon entropy. Shannon entropy, originally defined in statistical physics and information theory (Cover and Thomas 1991; Greiner et al. 1995), is used to measure the uncertainty removed or the information gained by performing an experiment. When it is applied to characterize DNA variation, entropy measures genetic diversity and extracts the maximal amount of information for a set of SNP markers (Hampe et al. 2003). The difference between cases and controls in entropy of the SNP markers is a measure of the association of the markers with diseases (Zhao et al. 2005). In this paper we propose a statistic based on entropy with high power for joint analysis for two-stage genome-wide association studies.

The main purpose of this article is to develop an entropy-based statistic with high power that is based on the nonlinear transformation of risk allele frequencies for joint analysis for two-stage genome-wide association studies. We compare the power among one stage analysis, linear joint analysis and entropy-based nonlinear joint analysis by simulation studies. To demonstrate that amplification of the differences in allele frequencies by a nonlinear test statistic will not cause false-positive problems, we investigate the type I error rates of the entropy-based nonlinear test statistic in a single-locus association test by simulations. Finally, we compare the power of linear joint analysis with that of entropy-based joint analysis when the same false discovery rate is controlled. From these results, we recommend we should use entropy-based joint analysis for genome-wide association studies.

Methods

We consider evaluating M markers using a case-control design, where the N cases and N controls are all unrelated individuals. Further, we assume that the markers are not in strong linkage disequilibrium with each other, hence the markers can be considered independent, and that the alleles in one locus are in Hardy–Weinberg equilibrium(HWE). We test every marker in a proportion (π_{samples}) of samples in stage 1 and select approximately $M \times \pi_{\text{markers}}$ markers for genotyping on the remaining $N \times (1 - \pi_{\text{samples}})$ cases and controls in stage 2.

Denote Z_1 and Z_2 to be the test statistics of a marker at stage 1 and stage 2, and C_1 and C_2 to be the critical values for stage 1 and stage 2, respectively. We denote $P_0(\cdot)$ and $P_A(\cdot)$ to be the probabilities of an event under the null and alterative hypotheses, respectively. Then from Satagopan et al. (2002, 2003, 2004), we know that the false-positive rate for a marker when using a two-stage strategy is

$$\alpha_{\text{markers}} = \alpha_{\text{genome}}/M = P_0(Z_1 > C_1)P_0(Z_2 > C_2|Z_1 > C_1),$$

where α_{genome} is any desired genome-wide false-positive rate (type I error rate). The power of the two-stage strategy is the probability of selecting a disease locus under the alternative hypothesis, which is

$$\begin{aligned}\text{Power} &= P_A(Z_1 > C_1, Z_2 > C_2)\\ &= P_A(Z_1 > C_1)P_A(Z_2 > C_2|Z_1 > C_1).\end{aligned}$$

Denote $\hat{P}_1^A$ and $\hat{P}_1^U$ to be the estimated risk allele frequencies in cases and controls in stage 1, respectively. The test statistic is defined in Skol et al. (2006)

$$Z_1 = \frac{\hat{P}_1^A - \hat{P}_1^U}{\sqrt{(\hat{P}_1^A(1-\hat{P}_1^A) + \hat{P}_1^U(1-\hat{P}_1^U))/(2N\pi_{\text{samples}})}}.$$

Under the null hypothesis of no association, and when a large number of samples $N\pi_{\text{samples}}$ is genotyped in stage 1, Z_1 follows a normal distribution with mean 0 and variance 1. We can determine a threshold C_1 for selecting markers for follow-up such that $P(|Z_1| > C_1) = \pi_{\text{markers}}$. Under the alternative hypothesis, the statistic Z_1 in large samples follows an approximate normal distribution with mean

$$\mu_1 = \frac{P^A - P^U}{\sqrt{(P^A(1-P^A) + P^U(1-P^U))/(2N\pi_{\text{samples}})}}$$

and variance 1, where P^A and P^U are the risk allele frequencies in cases and controls.

Entropy-based joint analysis

In statistical physics and information theory, entropy measures the uncertainty of random variables or the degree of non-structure within a system (Cover and Thomas 1991; Greiner et al. 1995). The entropy of a discrete variable or a system X is defined as:

$$S(X) = -\sum_i p(x_i)\log p(x_i),$$

where $p(x_i) = \text{Prob}(X = x_i)$. Entropy can be used to measure DNA variations at disease genes underlying a complex disease (Ackerman et al. 2003; Hampe et al. 2003; Zhao et al. 2005).

The entropies of risk allele at one marker in cases and controls are defined as $S^A = -p^A \log p^A$ and $S^U = -p^U \log p^U$, respectively.

Then the new entropy-based test statistic for an association test is defined as:

$$Z^e = \frac{\hat{S}^A - \hat{S}^U}{\sqrt{\frac{\hat{P}^A(1-\hat{P}^A)(1+\log\hat{P}^A)^2}{2N^A} + \frac{\hat{P}^U(1-\hat{P}^U)(1+\log\hat{P}^U)^2}{2N^U}}},$$

where $\hat{S}^A, \hat{S}^U, \hat{P}^A, \hat{P}^U$ are the estimators of S^A, S^U, P^A, and P^U, respectively.

From theorem 1.9 (Lehmann 1983), we know that the statistic Z^e is asymptotically distributed as a normal distribution with mean 0 and variance 1 under the null hypothesis of no association. Under the alterative hypothesis of association, Z^e is asymptotically distributed as a normal distribution with mean

$$\mu^e = \frac{S^A - S^U}{\sqrt{\frac{P^A(1-P^A)(1+\log P^A)^2}{2N^A} + \frac{P^U(1-P^U)(1+\log P^U)^2}{2N^U}}},$$

and variance 1.

In *stage 1*, the statistic for entropy-based joint analysis when $N^U = N^A = N$ is

$$Z_1^e = \frac{\sqrt{2N\pi_{\text{samples}}}(\hat{S}^A - \hat{S}^U)}{\sqrt{\hat{P}^A(1-\hat{P}^A)(1+\log\hat{P}^A)^2 + \hat{P}^U(1-\hat{P}^U)(1+\log\hat{P}^U)^2}}.$$

Under the null hypothesis of no association, and when a large number of samples $N\pi_{\text{samples}}$ is genotyped in stage 1, Z_1^e follows a normal distribution with mean 0 and variance 1. The threshold $C_1^e = \Phi^{-1}(1-\pi_{\text{markers}}/2)$ is determined for selecting markers for follow-up genotyping. So the probability that a marker will be selected for stage 2 genotyping is

$$P_1^e = 1 - \Phi(C_1^e - \mu_1^e) + \Phi(-C_1^e - \mu_1^e),$$

where $\mu_1^e = \frac{\sqrt{2N\pi_{samples}}(S^A - S^U)}{\sqrt{P^A(1-P^A)(1+\log P^A)^2 + P^U(1-P^U)(1+\log P^U)^2}}$. Similarly, an analogous statistic Z_2^e is calculated using genotype data only from stage 2.

In *stage 2*, the entropy-based statistic Z_{joint}^e is calculated using genotype data from both stages 1 and 2 as

$$Z_{joint}^e = \sqrt{\pi_{\text{samples}}}Z_1^e + \sqrt{1-\pi_{\text{samples}}}Z_2^e.$$

Conditional on the observed stage 1 statistic $Z_1^e = x$, the statistic for joint analysis Z_{joint}^e follows an approximate normal distribution in large samples with mean

$$\mu_{joint}^e = \frac{\sqrt{2N}(S^A - S^U)}{\sqrt{P^A(1-P^A)(1+\log P^A)^2 + P^U(1-P^U)(1+\log P^U)^2}} + \sqrt{\pi_{\text{samples}}}(x-\mu_1^e)$$

and variance $1 - \pi_{\text{samples}}$. Under the null hypothesis of no association, $\mu_{joint}^e = \sqrt{\pi_{\text{samples}}}x$. The critical value C_{joint}^e can be calculated iteratively by finding the threshold that satisfies $P_0(|Z_{joint}^e| > C_{joint}^e | |Z_1^e| > C_1^e) = \alpha_{\text{genome}}/(M\pi_{\text{markers}})$. The probability of detecting association in stage 2 in an entropy-based joint analysis is

$$\begin{aligned}
P_{joint}^e &= P_A(|Z_{joint}^e| > C_{joint}^e | |Z_1^e| > C_1^e) \\
&= \int_{C_1^e}^{\infty} P_A(|Z_{joint}^e| > C_{joint}^e | |Z_1^e| = x) f(x | |Z_1^e| > C_1^e)dx \\
&\quad + \int_{-\infty}^{-C_1^e} P_A\left(|Z_{joint}^e| > C_{joint}^e | |Z_1^e| = x\right) f\left(x | |Z_1^e| > C_1^e\right)dx \\
&= \int_{|x|>C_1^e}\int_{|y|>C_{joint}^e} \frac{1}{2\pi\sqrt{1-\pi_{\text{samples}}}[1-\Phi(C_1^e-\mu_1^e)+\Phi(-C_1^e-\mu_1^e)]} \\
&\quad \exp\left(-\frac{(y-\mu_0^e)^2 - 2\pi_{\text{samples}}(x-\mu_1^e)(y-\mu_0^e) + (x-\mu_1^e)^2}{2(1-\pi_{\text{samples}})}\right)dxdy,
\end{aligned}$$

where

$$\mu_0^e = \frac{\sqrt{2N}(S^A - S^U)}{\sqrt{P^A(1-P^A)(1+\log P^A)^2 + P^U(1-P^U)(1+\log P^U)^2}}.$$

The power of entropy-based joint analysis for two-stage genome-wide association studies is

$$\int_{|x|>C_1^e} \int_{|y|>C_{joint}^e} \frac{1}{2\pi\sqrt{1-\pi_{\text{samples}}}} \exp\left(-\frac{(y-\mu_0^e)^2 - 2\pi_{\text{samples}}(x-\mu_1^e)(y-\mu_0^e) + (x-\mu_1^e)^2}{2(1-\pi_{\text{samples}})}\right) dxdy.$$

Disease models

Denote genotype relative risks (GRRs) to be R_1 and R_2, the disease prevalence to be $Prev$, and the risk allele frequency to be P_d. Then the disease penetrances $f_i = P(\text{affected}|i \text{ copies of } d \text{ allele})$ $(0 \le i \le 2)$ are

$$f_0 = Prev/((1-P_d)^2 + 2P_d(1-P_d)R_1 + P_d^2R_2),$$
$$f_1 = f_0R_1,\ f_2 = f_0R_2.$$

Therefore

$$P(dd|\text{Cases}) = \frac{f_2P_d^2}{Prev},$$

$$P(Dd|\text{Cases}) = \frac{2f_1(1-P_d)P_d}{Prev},$$

$$P(DD|\text{Cases}) = \frac{f_0(1-P_d)^2}{Prev};$$

$$P(dd|\text{Controls}) = \frac{(1-f_2)P_d^2}{1-Prev},$$

$$P(Dd|\text{Controls}) = \frac{2(1-f_1)(1-P_d)P_d}{1-Prev},$$

$$P(DD|\text{Controls}) = \frac{(1-f_0)(1-P_d)^2}{1-Prev}.$$

So, risk allele frequencies at a marker locus in cases and controls can be achieved respectively as

$$\frac{f_2P_d^2}{Prev} + \frac{f_1(1-P_d)P_d}{Prev} \text{ and } \frac{(1-f_2)P_d^2}{1-Prev} + \frac{(1-f_1)(1-P_d)P_d}{1-Prev}.$$

Results

In order to apply the entropy-based statistic to genome-wide association studies, we first examined the property of this test statistic in the simple case, single-locus case-control association studies. In the methods, we have shown that when the sample size is large enough to apply large-sample theory, the distribution of the entropy-based statistic under the null hypothesis of no association is asymptotically a normal distribution. To examine whether the asymptotic result of the entropy-based test statistic still holds for a small sample size, 200 individuals were randomly generated. A total of 10,000 simulations were performed. In each simulation, we calculated the entropy-based test statistic Z^e.

Table 1 summarizes the estimated type I error rates of the test statistic Z^e for sample sizes from 100 to 500 individuals for association test. It shows that the estimated type I error rates of the test statistic Z^e are not appreciably different from the nominal levels $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.005$. Table 2 summarizes the power of the entropy-based statistic in single-locus association studies for sample sizes from 100 to 500 individuals, using a multiplicative model with $R_1 = 1.60$ and $R_2 = 2.56$ and disease prevalence of 0.10. It shows that the power of entropy-based test is higher than that using a linear function of risk allele frequency.

Table 1 Evaluated type I error rates for the test statistic Z^e in single-locus association test (10,000 simulations)

Sample size	Type I error rates for nominal level		
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$
100	0.0500	0.0113	0.0062
150	0.0493	0.0090	0.0053
200	0.0547	0.0103	0.0059
250	0.0484	0.0109	0.0058
300	0.0488	0.0106	0.0050
350	0.0511	0.0099	0.0043
400	0.0516	0.0120	0.0063
450	0.0520	0.0119	0.0062
500	0.0492	0.0110	0.0050

Table 2 Power of the test statistic Z^e in single-locus association study (10,000 simulations)

$P^A - P^U$	2%	3%	4%	5%	
100	LF	0.1793	0.2492	0.3021	0.3456
	Entropy	0.1941	0.2538	0.3180	0.3472
300	LF	0.4236	0.5858	0.7042	0.8010
	Entropy	0.4241	0.6106	0.7184	0.8066
500	LF	0.6404	0.8090	0.9035	0.9531
	Entropy	0.6424	0.8180	0.9086	0.9553

Uses a multiplicative model with $R_1 = 1.60$ and $R_2 = 2.56$ and risk allele frequencies $P_d \leq 0.10$ and disease prevalence of 0.10. The significant level is 0.05

Now we apply the new statistic to joint analysis for two-stage genome-wide association studies. First, we compare the powers of one-stage, linear joint analysis and entropy-based nonlinear joint analysis at $\alpha_{\text{genome}} = 0.05$ for a wide range of proportions (π_{samples}) of samples in stage 1, proportions (π_{markers}) of markers selected for follow-up genotyped in stage 2 and four different genetic models with risk allele frequencies 0.05 and 0.10. All the results show that entropy-based nonlinear joint analysis is more powerful and a more efficient design for genome-wide association studies (Figs. 1, 2).

We then investigate the power of nonlinear joint analysis as a function both of frequencies of risk allele under multiplicative genetic models (Fig. 3) and of proportions (π_{markers}) of markers selected for follow-up detection in stage 2 (Fig. 4). We find that the power of the entropy-based joint analysis is always higher than linear joint analysis when the frequency of risk allele is small. However, as the frequency of risk allele increases, the powers of these two joint analyses are comparable.

We also investigate the samples sizes needed to detect the genetic variants with different effect sizes (Fig. 5) by linear and entropy-based joint analyses in two-stage design. The sample size needed for entropy-based joint analysis is

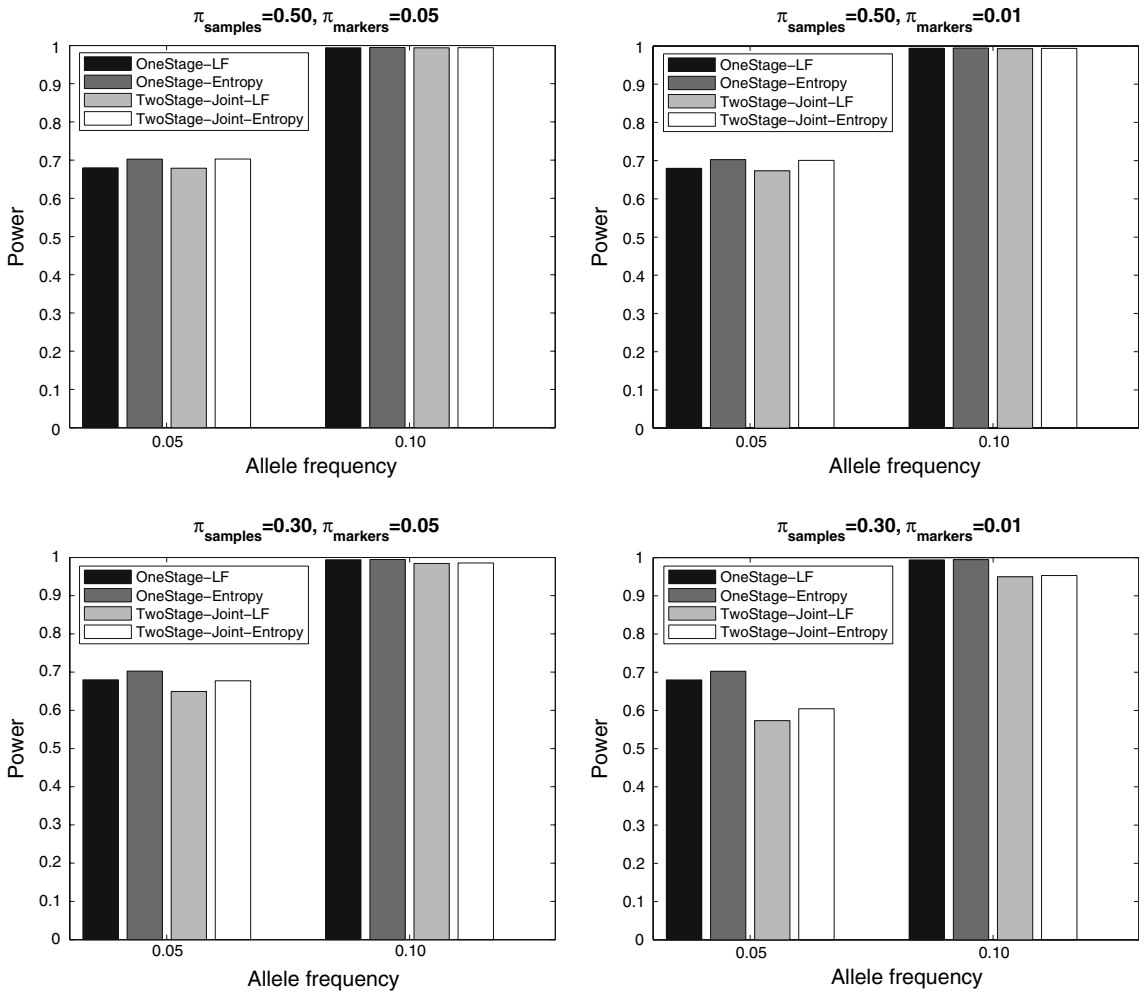

Fig. 1 Power of linear and entropy-based joint analyses with 2,000 cases and 2,000 controls genotyped on 300,000 independent markers with $\alpha_{\text{genome}} = 0.05$. Using a multiplicative genetic model with $R_1 = 1.40$ and $R_2 = 1.96$ and disease prevalence of 0.10

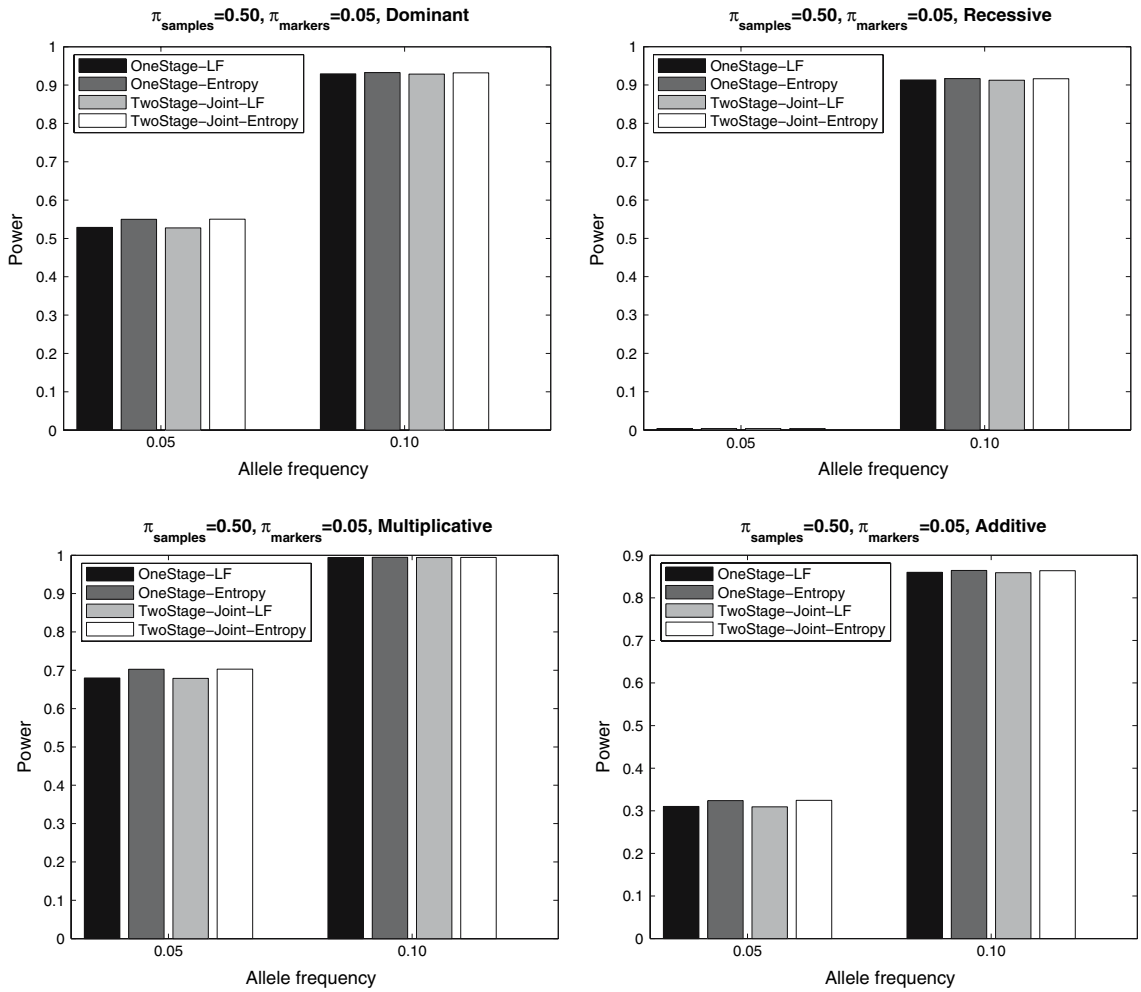

Fig. 2 Power of linear and entropy-based joint analyses with 2,000 cases and 2,000 controls genotyped on 300,000 independent markers with α_{genome} = 0.05 under four different genetic models and disease prevalence of 0.10. Dominant model: R_1 = 1.60, R_2 = 1.60; recessive model: R_1 = 1, R_2 = 6; multiplicative model: R_1 = 1.60, R_2 = 2.56; additive model: R_1 = 1.50, R_2 = 2

less than that needed for linear joint analysis to get the same power. For obtaining power of 80%, for the genetic variants P_d = 0.10 with modest and large effect sizes at a multiplicative model, we suppose the sample size are respectively to be 2,540 and 1,227 for genetic variants with effect sizes GRR = 1.4 and 1.6, respectively.

When controlling the false discovery rate, we compare the power of linear and entropy-based joint analyses in two-stage genome-wide association studies as a function of the difference of risk allele frequencies between cases and controls (Fig. 6). The power of entropy-based joint analysis is higher than that of the linear joint analysis controlling the same false discovery rate when detecting the genetic variants with a small frequency. It makes sense if we want to attain the same power for two joint analyses, then the false positive rate of linear joint analysis will increase. For example, the false-positive rate increases from 0.05 to nearly 0.10 when the same power in two joint analyses is achieved for $\pi_{\text{samples}} = 0.30, \pi = 0.01$, GRR = 1.60, and $P^A - P^U = 0.04$.

In Table 3, we compare the power of entropy-based joint analysis with that of the linear joint analysis under four different genetic models. We find that the powers of entropy-based joint analysis are 2% higher than that of the linear joint analysis under the risk allele frequency of 0.05 by simulations. In Table 4, we evaluate the sample size needed in entropy-based joint analysis, and it shows that there are fewer samples needed in entropy-based joint analysis than that needed in linear joint analysis. In Table 5, we compare the power of linear and entropy-based joint analyses when controlling the false discovery rate for the fixed allele frequency difference between cases and controls. We can find that the false discovery rate of linear joint analysis increases from 0.05 to 0.1 when getting a power of 0.93 compared with the entropy-based joint analysis. All results show that the entropy-based analysis is

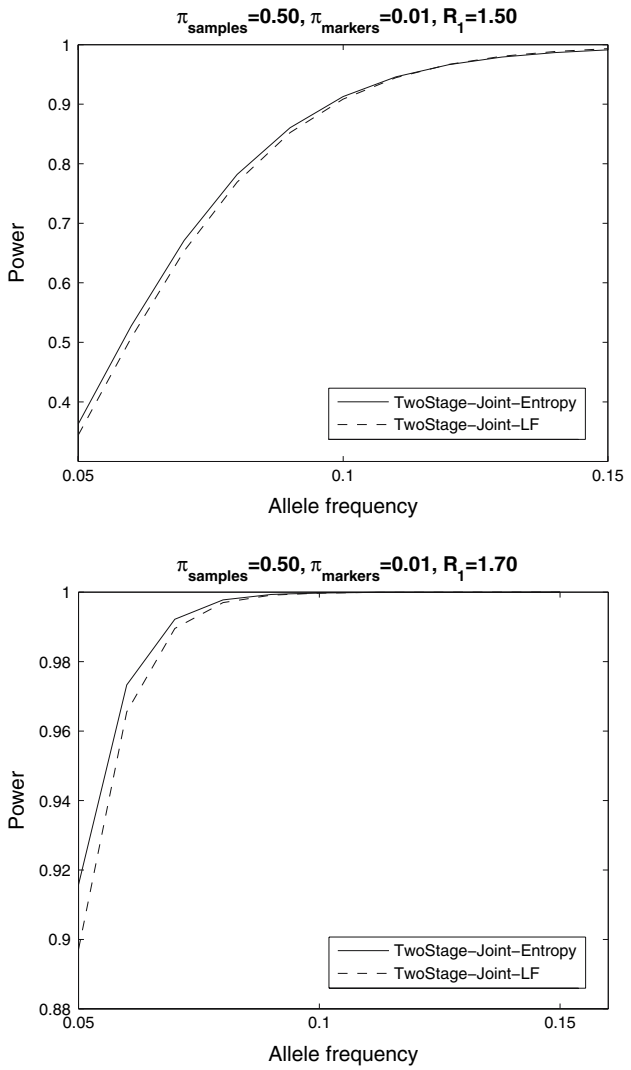

Fig. 3 Power of linear and entropy-based joint analyses as a function of the frequencies of risk allele with 2,000 cases and 2,000 controls genotyped on 300,000 markers with α_{genome} = 0.05; it uses two multiplicative genetic models (R_1 =1.50, R_2 = 2.25 and R_1 = 1.70, R_2 = 2.89) and disease prevalence of 0.10

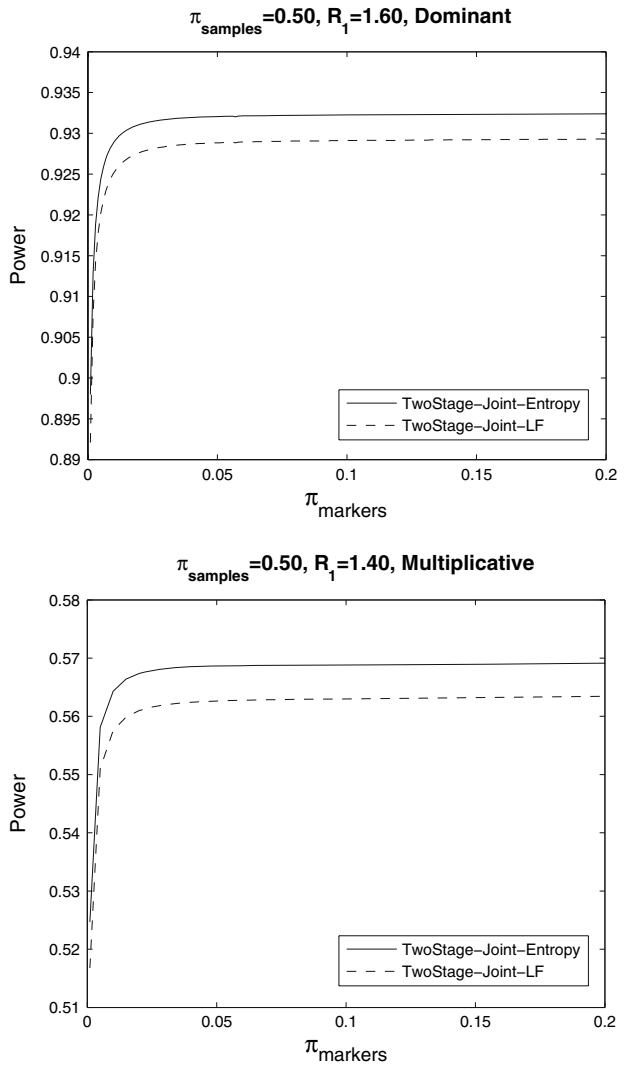

Fig. 4 Power of linear and entropy-based joint analyses as a function of $\pi_{markers}$ with 2,000 cases and 2,000 controls genotyped on 300,000 markers with α_{genome} = 0.05; it uses dominant (R_1 = 1.60, R_2 = 1.60) and multiplicative (R_1 = 1.40, R_2 = 1.96) genetic models with disease prevalence of 0.10 and risk allele frequency of 0.10

more powerful and needs fewer samples for attaining the same power and achieving the same false discovery rate. These make sense, as entropy-based joint analysis uses a nonlinear function of risk allele frequencies so that it makes full use of data information from all samples.

Discussion

We have shown that the entropy-based joint analysis for two-stage genome-wide association design is a more efficient and more powerful strategy to identify genetic variants with variant effect sizes associated with a disease when testing a large number of markers using unrelated case-control samples. For achieving an overall power of 90% when detecting genetic variants both with small frequency and with small to large effects, the sample size needed in entropy-based joint analysis is about 30 fewer than that needed in linear joint analysis.

Genome-wide disease-association mapping has been herald as the study design of the next generation (Marchini et al. 2005); two-stage designs have been a promising strategy for genome-wide association studies, but the lack of analytical methods to use genotype data fully and sufficiently is a large stumbling block (Lin et al. 2004). So, we should commit ourselves to find more powerful and more efficient methods (or statistics) in the near future. The traditional test statistic in Skol et al. (2006) is a linear

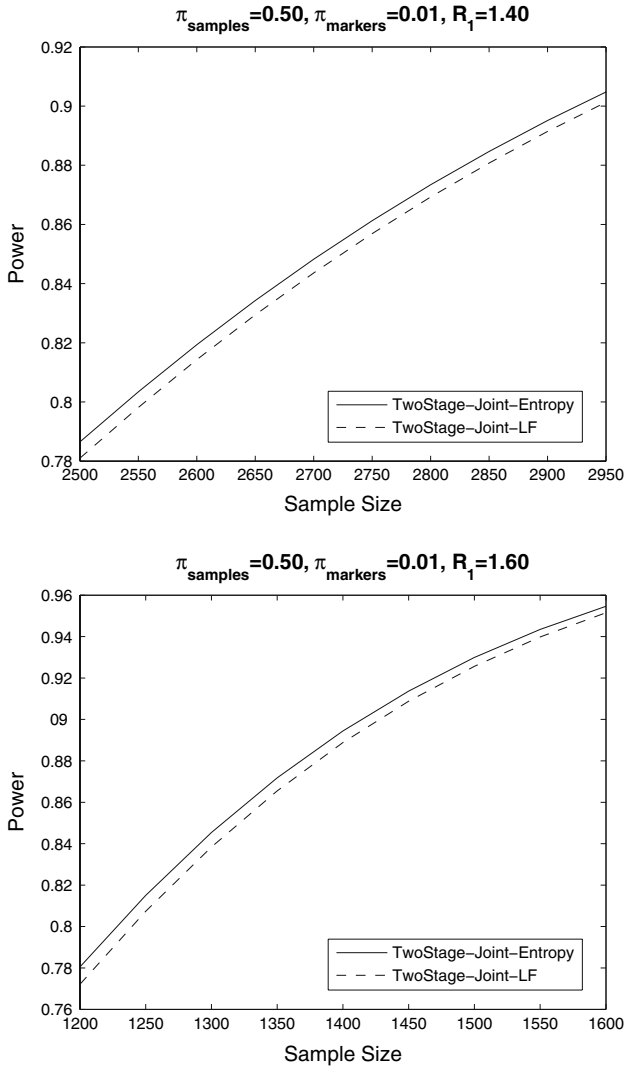

Fig. 5 Power of linear and entropy-based joint analyses with variant sample sizes on 300,000 independent markers with $\alpha_{\text{genome}} = 0.05$. It uses two multiplicative models with $R_1 = 1.40$, $R_2 = 1.96$ and $R_1 = 1.60$, $R_2 = 2.56$ respectively and *prevalence* = 0.10, risk allele frequency of 0.10

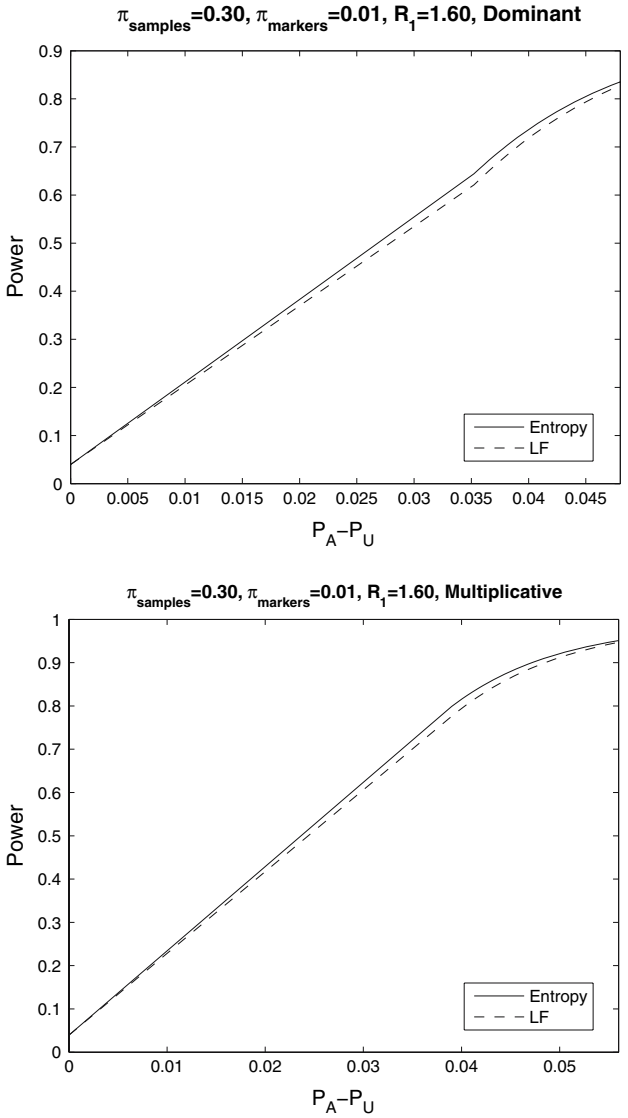

Fig. 6 Power of linear and entropy-based joint analyses under controlling the same false discovery rate, with sample size 2,000 cases and 2,000 controls on 300,000 independent markers. It uses dominant ($R_1 = 1.60, R_2 = 1.60$) and multiplicative ($R_1 = 1.60, R_2 = 2.56$) models, *prevalence* = 0.10, $\pi_{\text{samples}} = 0.30$ and $\pi_{\text{markers}} = 0.01$

function of $(P^A - P^U)$ in risk allele frequencies between cases and controls. Here, we introduce a nonlinear function of risk allele frequencies in cases and controls, entropy, $(S^A - S^U)$ to develop novel test statistics with high power for detecting the genetic variants underlying the disease.

We investigate the distribution of a nonlinear entropy-based statistic under the null hypothesis by simulation studies. To validate the test statistic, we calculate the type I error rates of the entropy-based statistic by simulations. It shows that the type I error rates of the entropy-based statistic are close to the nominal significance levels. To evaluate the performance of the entropy-based joint analysis, we compare the power of the entropy-based statistic in single-locus association study with that of the statistic using linear function of risk allele frequencies in cases and controls by simulations. The results show that the entropy-based statistic has a higher power than the statistic using the linear function of risk allele frequencies in cases and controls. However, since the power of the statistic is a complex issue, there is not one statistic that is uniformly more powerful (Zhao et al. 2006). The entropy-based analysis is also not more powerful in all situations. When a large difference of rare risk allele frequencies between cases and controls appears, that is, $|P^A - P^U| > 0.07$, the linear joint analysis is more powerful than the entropy-based joint analysis when detecting rare genetic variants

Table 3 Power of entropy-based joint analyses for two-stage genome-wide association studies under four genetic models

Models	R_1	R_2	$\pi_{samples}$	$\pi_{markers}$	Power	
					LF	Entropy
Dominant	1.5	1.5	1	0.05	0.62	0.64
			0.5		0.62	0.64
			0.3		0.59	0.61
Recessive	1	11	1	0.05	0.76	0.77
			0.5		0.76	0.77
			0.3		0.73	0.75
Multiplicative	1.5	2.56	1	0.05	0.81	0.82
			0.5		0.80	0.82
			0.3		0.77	0.79
Additive	1.5	2	1	0.05	0.72	0.73
			0.5		0.71	0.73
			0.3		0.68	0.70

There are 3,000 cases and 3,000 controls genotyped on 300,000 independent markers, the significance level of two-stage genome-wide design is 0.05 and uses a risk allele frequency of 0.05 and disease prevalence of 0.10

R_1 = *P(Affected\|Dd)/P(Affected\|DD)*, R_2 = *P(Affected\|dd)/P(Affected\|DD)*; LF, represents linear joint analysis based on a linear function of risk allele frequencies in cases and controls (all the "LF" in the following have the same meaning as this one and we will omit this)

Table 4 Sample size to attain the desired significance level of two-stage genome-wide design 0.05 and power of 80% for various rare allele frequency differences and population allele frequencies

$P^A - P^U$ (%)	R_1 = 1.30		R_1 = 1.40		R_1 = 1.60	
	LF	Entropy	LF	Entropy	LF	Entropy
3	4,337	4,319	3,285	3,248	2,331	2,269
4	–	–	2,340	2,324	1,754	1,713
5	–	–	–	–	1,410	1,385

Where all samples are genotyped on 300,000 independent markers, $\pi_{samples} = 0.50, \pi_{markers} = 0.01$, and use a multiplicative model ($R_2 = R_1^2$) with risk allele frequencies $P_d \leq 0.10$ and with disease prevalence of 0.10

– Means that such differences in rare risk allele frequencies between cases and controls are practically impossible to appear in a real world

with variant genetic effects. However, these differences in rare risk allele frequencies between cases and controls are practically unrealistic in real-world studies of rare variants/common diseases.

Subsequently, we apply the entropy-based statistic to two-stage genome-wide association studies. We compare the power of entropy-based nonlinear joint analysis with that of the linear joint analysis by simulations. The results show that the power of the entropy-based joint analysis is higher than the power of the linear joint analysis in most cases when detecting rare genetic variants with variant

Table 5 Power of entropy-based joint analyses for two-stage genome-wide association studies when controlling FDR

$P^A - P^U$ (%)	FDR = 0.10		FDR = 0. 05		FDR = 0.01	
	LF	Entropy	LF	Entropy	LF	Entropy
2	0.22	0.24	0.19	0.21	0.13	0.14
3	0.60	0.63	0.57	0.60	0.50	0.53
4	0.83	0.85	0.82	0.84	0.79	0.81
5	0.93	0.93	0.92	0.93	0.92	0.92

Where 2,000 cases and controls are genotyped on 300,000 independent markers, $\pi_{samples} = 0.30, \pi_{markers} = 0.01,$. It uses a multiplicative model with R_1 = 1.40 and R_2 = 1.96 and a disease prevalence of 0.10

FDR false discovery rate

genetic effects. However, entropy is one of the nonlinear transformations of risk allele frequencies between cases and controls. The general forms of nonlinear transformations $f(P^A, P^U)$ of risk allele frequencies in cases and controls should be investigated in the future.

Here we have described entropy-based joint analysis for two-stage genome-wide association studies using independent genetic markers. But this assumption will be violated when some markers are in linkage disequilibrium. For two genetic variants each with a small effect, we should consider the interaction between loci in genome-wide association studies when they contribute modest or large effects in combination. This will be an inevitable and promising field for genome-wide association studies.

The simulations show that for a given sample size, we should genotype half of the individuals on all markers in the first stage and select the 5% of markers for follow-up genotyping in the second stage using the entropy-based statistic, which provides a practical cost-effective strategy to search for rare genetic variants in association studies. The simulations also show that for searching for rare genetic variants with moderate effects (R_1 = 1.4, 1.6), the sample size is approximately 2,000 for the fixed rare allele frequency difference (4%, 5%) by using the entropy-based joint analysis.

In multiple tests, there is an increasing trend to use a false discovery rate as a measure of global error instead of using overall type I error rate. This article compares the power of entropy-based joint analysis with that of linear joint analysis controlling the false discovery rate when its level is set to be the same, which is usually done in the literature (Benjamini and Hochberg 1995; Zou and Zuo 2006; Zuo et al. 2006). The results also show that entropy-based joint analysis leads to higher power than linear joint analysis when controlling the same false discovery rate, which makes sense.

In conclusion, numerous genome-wide association studies for a range of diseases are being planned or are already underway. Developing new statistical methods that

can deal with such large-scale studies is urgently needed to explore the etiology of complex diseases. Two-stage designs are more efficient and powerful, comparable to the one stage design. The results in this paper show that the entropy-based joint analyses are more powerful and need fewer samples for attaining the same power and achieving the same false discovery rate. Therefore, we suggest that we should use entropy-based joint analysis for two-stage genome-wide association studies.

Acknowledgments We would like to thank two referees for very helpful comments on an earlier draft. This work was supported by grant DMS 0234078 from the National Science Foundation to Y. Zuo.

References

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Tayor T, Ward R, Molyneux M, Pinder M et al (2003) Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Bio 4(4):R24.10

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57:289–300

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York, pp 12–15

Greiner W, Neise L, Stocker H (Translator) (1995) Thermodynamics and statistical mechanics. Springer, New York, pp 121–135

Hampe J, Schreiber S, Krawczak M (2003) Entropy-based SNP selection for genetic association studies. Hum Genet 114:36–43

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL,Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

Lehmann EL (1983) Theory of point estimation. Wiley, New York, pp 343–344

Lin DY (2006) Evaluating statistical significance in two-stage genomewide association studies. Am J Hum Genet 78:505–509

Lin S, Chakravarti A, Culter DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36:1181–1188

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. Biometrics 58:163–170

Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene- disease association studies with sample size contraints. Biometrics 60:589–597

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213

Thomas D, Xie RR, Gebregzibher M (2004) Sampling designs for gene association studies. Genet Epidemiol 27:401–414

Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345

Zhao JY, Boerwinkle E, Xiong MM (2005) An entropy-based statistic for genomewide association studies. Am J Hum Genet 77:27–40

Zhao JY, Jin L, Xiong MM (2006) Nonlinear tests for genome-wide association studies. Genetics 174:1529–1538

Zou GH, Zuo YJ (2006) On the sample size requirement in genetic association tests when the proporting of false positives is controlled. Genetics 172:687–691

Zuo YJ, Zou GH, Zhao HY (2006) Two-stage designs in case-control association analysis. Genetics 173:1747–1760