Statistica Sinica **19** (2009), 1769-1786

TWO-STAGE GENOME-WIDE ASSOCIATION STUDIES WITH DNA POOLING AND GENETIC MODEL SELECTION

Min Yuan[1], Yaning Yang[1] and Gang Zheng[2]

[1] *University of Science and Technology of China*
and [2] *National Heart, Lung and Blood Institute*

Abstract: The two-stage design is a common cost-effective approach for genome-wide association studies. The first stage serves as a screening to identify a subset of single-nucleotide polymorphisms (SNPs) from 100,000 to 500,000 SNPs using a fraction of case-control samples. In the second stage, only the selected SNPs are genotyped using the remaining case-control samples. On the other hand, DNA pooling is another common strategy to save genotyping cost. In this article, we propose a method using DNA pooling in the first stage and genotype-based analysis in the second stage. A joint analysis to combine both stages is applied to a two-stage design with DNA pooling when the underlying genetic model is known. When the genetic model is unknown, we use a robust procedure in the joint analysis by applying genetic model selection in the second stage based on the difference of Hardy-Weinberg disequilibrium coefficients between cases and controls. Performance of our method and comparison with other approaches are investigated by simulation studies.

Key words and phrases: Cost-effective design, DNA pooling, genetic model selection, joint analysis, robustness, trend tests, two-stage.

1. Introduction

In candidate-gene association studies, one tests association between a disease and the candidate genetic marker. Since hundreds of thousands of single-nucleotide polymorphisms (SNPs) can now be genotyped, genome-wide association study (GWAS) becomes a promising and powerful approach to identify true association between genetic markers and complex diseases. Although genotyping costs have been reduced recently, cost-effective designs for GWAS are still desirable. Various two-stage designs have been proposed recently (see e.g., Satagopan, Verbel, Venkatraman, Offit and Begg (2002), Satagopan and Elston (2003), Satagopan, Venkatraman and Begg (2004), Thomas, Xie and Gebregziabher (2004), Thomas, Haile and Duggan (2005), Lin (2006), Wang, Thomas, Pe'er and Stram (2006), Skol, Scott, Abecasis and Boehnke (2006), Zuo, Zou and Zhao (2006), Bukszar and van den Oord (2006), Ji, Stephen, Chad, Nancy and Derek (2007),

and Dube, Schmidt and Hauser (2007)). One common feature of these two-stage designs is that a fraction of samples are genotyped for all SNPs in a first stage. An association test is then applied to one SNP at a time. The most significant SNPs are selected and then genotyped for the remaining samples. Association analysis is then conducted for the selected SNPs in a second stage conditional on the results in the first stage (Elston, Lin and Zheng (2007)). After a small fraction of SNPs is identified by the above two-stage scan, more powerful and focused analysis can be conducted, e.g., haplotype analysis, multi-marker analysis, fine mapping, and replication (Hoh and Ott (2003), Marchini, Donnelly and Cardon (2005), Schaid, McDonnell, Hebbring, Cunningham and Thibodeau (2005), and Wang, Zhu and Elston (2007)). Most research papers focus on cost-effective two-stage designs for GWAS. In this article, however, we do not consider the cost-effectiveness but focus on some analysis strategies for a given design (e.g., given the proportion of samples used and percentage of SNPs selected in each stage).

DNA pooling is another cost-effective technique (Barcellos, Klitz, Field, Tobias, Bowcock, Wilson, Nelson, Nagatomi and Thomson (1997), Sham, Bader, Craig, O'Donovan and Owen (2002), and Norton, Williams, O'Donovan and Owen (2004)) in which several pools of DNA are allelotyped rather than each individual being genotyped. Zuo et al. (2006) applied the DNA pooling to the first stage of a two-stage design. In their second stage, each individual of the remaining samples is genotyped for the selected SNPs. In Skol et al. (2006), individuals are genotyped in both stages. Thus, the design of Zuo et al. (2006) would save more genotyping cost than that of Skol et al. (2006). For the analysis, Zuo et al. (2006) combined case-control data in the two stages into a single case-control sample and applied a single allele-based test (ABT) statistic. On the other hand, Skol et al. (2006) considered a joint analysis by weighting the two ABTs from the two stages with weights proportional to sample sizes in the two stages. One advantage of using the joint analysis is that it allows different allele frequencies in samples (heterogeneity) from the two stages. When the ABT is used, ignoring possible measurement errors, application of DNA pooling with a joint analysis would reduce more genotyping cost while retaining the same statistical power compared to individual genotyping with a joint analysis.

The Cochran-Armitage trend test (CATT) is proposed for analysis of ordered case-control data (Armitage (1955), Cochran (1954), and Sasieni (1997)). Optimal CATTs are available for different genetic models (Sasieni (1997) and Freidlin, Zheng, Li and Gastwirth (2002)). We integrate the DNA pooling of Zuo et al. (2006) with the joint analysis of Skol et al. (2006) to examine the power gain while the optimal CATT and the ABT-based two-stage strategies are employed. This, however, requires us to know the genetic model. When the genetic model is unknown, which is usually the case in practice, we propose a robust

joint analysis with genetic model selections followed by using the corresponding optimal CATT in the second stage, while DNA pooling technique is used in the first stage. Numerical and simulation results are presented to compare power and robustness of our method with the existing procedures.

2. Background

2.1. Notation, genetic models and association tests

Consider a SNP with alleles A and a and frequency $P(A) = p$. Denote the three genotypes by $g_0 = aa$, $g_1 = Aa$ and $g_2 = AA$, the disease prevalence by $K = P(\text{case})$, and the penetrance by $f_l = P(\text{case}|g_l)$ for $l = 0, 1, 2$. For a case-control study with r cases and s controls, let x_i and y_j be, respectively, the number of allele A for the ith case and the jth control for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. Write $p_l = P(x_i = l)$ and $q_l = P(y_j = l)$ for $l = 0, 1, 2$. The null hypothesis is $H_0 : p_l = q_l = P(g_l)$. Genotype counts are r_l in cases and s_l in controls for g_l, $l = 0, 1, 2$. Then $r_l = \sum_{i=1}^{r} I(x_i = l)$ and $s_l = \sum_{j=1}^{s} I(y_j = l)$, where $I(\cdot)$ is the indicator function. The counts (r_0, r_1, r_2) and (s_0, s_1, s_2) follow multinomial distributions $Mul(r, (p_0, p_1, p_2))$ and $Mul(s, (q_0, q_1, q_2))$, respectively. Denote the margins by $n_l = r_l + s_l$ and the total sample size by $n = r + s$.

Denote genotype relative risks (GRRs) by $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$ ($f_0 > 0$). We assume that A is the risk allele and that risk increases with the number of allele A in the genotype, i.e., $\lambda_2 \geq \lambda_1 \geq 1$. Four commonly used genetic models are recessive (REC), additive (ADD), multiplicative (MUL), and dominant (DOM), corresponding to $\lambda_1 = 1$, $\lambda_1 = (\lambda_2 + 1)/2$, $\lambda_2 = \lambda_1^2$ and $\lambda_2 = \lambda_1$, respectively.

Two common association tests are ABT and CATT (Sasieni (1997)). The ABT compares the frequencies of allele A in cases and controls, while the CATT compares the genotype distributions in cases and controls. Three CATTs are available depending on the genetic models. The same CATT is used for ADD or MUL (Freidlin et al. (2002) and Zheng, Freidlin, Li and Gastwirth (2003)). When Hardy-Weinberg equilibrium (HWE) holds in the combined case-control samples, the ABT and the additive CATT (optimal for the ADD model) are asymptotically equivalent (Sasieni (1997)). The ABT (T_{ABT}) and CATT (T_θ) are given by

$$T_{\text{ABT}} = \frac{(\hat{p}_1/2 + \hat{p}_2) - (\hat{q}_1/2 + \hat{q}_2)}{\{\hat{p}(1-\hat{p})(1/(2r) + 1/(2s))\}^{1/2}}, \tag{2.1}$$

$$T_\theta = \frac{(\hat{p}_2 + \theta\hat{p}_1) - (\hat{q}_2 + \theta\hat{q}_1)}{[\{(\hat{p}_2 + \theta^2\hat{p}_1) - (\hat{p}_2 + \theta\hat{p}_1)^2\}(1/r + 1/s)]^{1/2}}, \tag{2.2}$$

where $\hat{p}_l = r_l/r$, $\hat{q}_l = s_l/s$, $\hat{p} = n_l/n$, and $\theta = 0, 1/2, 1$ for the REC, ADD/MUL and DOM models. Under H_0, both tests are asymptotically $N(0,1)$.

2.2. Genetic model selections

When the true genetic model is unknown, T_θ cannot be directly used. The genetic model, however, may be detected using Hardy-Weinberg disequilibrium (HWD) coefficient, denoted by $\delta = P(AA) - \{P(AA) + P(Aa)/2\}^2$. Zaykin and Nielsen (2000) and Song and Elston (2006) applied the difference of HWD in cases and controls for testing association. Denote the HWD coefficients in cases and controls by δ_1 and δ_0. The HWD trend test (Song and Elston (2006)) can be written as $T_{\text{HWD}} = (rs/n)^{1/2}(\hat{\delta}_1 - \hat{\delta}_0)/[\{1 - n_2/n - n_1/(2n)\}\{n_2/n + n_1/(2n)\}]$, which asymptotically follows $N(0,1)$ under H_0.

Wittke-Thompson, Pluzhnikov and Cox (2005), Suh and Li (2007) and Zheng and Ng (2008) studied the relationship between genetic models and HWD. Zheng and Ng (2008) showed that, when HWE holds in the population, $\delta_1 > \delta_0$ under the REC model and $\delta_1 < \delta_0$ under DOM model, regardless of the risk allele. Thus, they used $T_{1/2}$ to test association unless $T_{\text{HWD}} > c_0$, under which they selected the REC model, and used T_0, or $T_{\text{HWD}} < -c_0$, under which they selected the DOM model and used T_1, where $c_0 = 1.645$ was used. This approach was referred to as genetic model selection (GMS), which is more robust than some existing methods and also robust to departure from HWE (Zheng and Ng (2008)).

3. Two-stage Design with DNA Pooling and Joint Analysis

Here we integrate the DNA pooling and the joint analysis of Skol et al. (2006) into a two-stage design. Due to DNA pooling, the ABT is the only test that can be used for the first stage. In the second stage, we could use the ABT as did in Skol et al. (2006), the optimal CATT when the genetic model is known, or the GMS when the model is unknown.

Similar to Zuo et al. (2006), in addition to r cases and s controls allelotyped in stage 1 with DNA pooling, an additional r_* cases and s_* controls are individually genotyped in stage 2 for the selected SNPs. In stage 1, cases and controls are grouped into m pools and the numbers of cases and controls in each pool are h_1 and h_0, respectively ($r = mh_1$ and $s = mh_0$). We assume a simple pooling measurement error mechanism (Barratt, Payne, Rance, Nutland, Todd and Clayton (2002)) that assumes the estimated allele frequencies from the pooled samples is equal to the true frequencies in the samples plus a disturbance variable that is $N(0, \epsilon^2)$. Usually, ϵ^2 needs to be estimated using the replicates of the DNA pooling from other sources (existing pooled data or prior knowledge). Here, however, we assume ϵ^2 is known, because it can be estimated in practice during the genotyping process with a given genotyping platform (Barratt et al. (2002)).

3.1. Using the ABTs in both stages

The ABT for pooled data can be written as

$$T_{\text{pool}} = \frac{\hat{p}_1^{\text{pool}} - \hat{p}_0^{\text{pool}}}{[2\epsilon^2/m + \hat{p}^{\text{pool}}(1-\hat{p}^{\text{pool}})\{1/(2r) + 1/(2s)\}]^{1/2}},$$

where $\hat{p}_0^{\text{pool}}$, $\hat{p}_1^{\text{pool}}$ and $\hat{p}^{\text{pool}}$ are the estimates of allele frequency in controls, cases, and combined samples (details are given in Appendix A). Under H_0, T_{pool} is asymptotically $N(0,1)$. For the second stage with additional r_* cases and s_* controls, we denote the ABT test as T_{ABT}. Denote the sample proportion in the first stage as $\omega = n/(n+n_*)$ and $n_* = r_* + s_*$. Following the joint analysis method of Skol et al. (2006), we propose the following joint test

$$J_{\text{ABT}} = \omega^{1/2} T_{\text{pool}} + (1-\omega)^{1/2} T_{\text{ABT}}. \tag{3.1}$$

The test statistic in (3.1) combines the design of Zuo et al. (2006) with DNA pooling in stage 1 and the joint analysis of Skol et al. (2006) in stage 2. To apply J_{ABT} with a total of M SNPs, we assume a fraction of $100\alpha_1\%$ top-ranked SNPs are selected in stage 1. Then, following Skol et al. (2006), to control the genome-wide level at α, we need to determine thresholds c_1 and c_2 such that, assuming A is the risk allele after stage 1 analysis,

$$P_{H_0}\Big(|T_{\text{pool}}| > c_1\Big) = \alpha_1, \tag{3.2}$$

$$P_{H_0}\Big(|T_{\text{pool}}| > c_1,\ |J_{\text{ABT}}| > c_2,\ T_{\text{pool}} \cdot T_{\text{ABT}} > 0\Big) = \frac{\alpha}{M}. \tag{3.3}$$

The two ABTs have the same sign because the same risk allele is identified. The formula for calculating c_2 and asymptotic power derived by Skol et al. (2006) can be applied, but the asymptotic covariances of the statistics T_{pool} and J_{ABT} under H_0 and a specific alternative H_1 are different because of DNA pooling (Appendix A). The asymptotic power of the joint analysis J_{ABT} can be written as (3.3), but evaluated under H_1 (see Appendix A).

3.2. Using the ABT in stage 1 and optimal CATT in stage 2

Because of the DNA pooling, the T_{pool} is the only statistic to use in stage 1. In stage 2, since individual genotypes are obtained, the CATT (2.2) can be calculated. Therefore, we modify J_{ABT} in (3.1) as

$$J_\theta = \omega^{1/2} T_{\text{pool}} + (1-\omega)^{1/2} T_\theta, \tag{3.4}$$

where θ is chosen based on the known genetic model. Accordingly, (3.3) becomes

$$P_{H_0}\Big(|T_{\text{pool}}| > c_1,\ |J_\theta| > c_2^*,\ T_{\text{pool}} \cdot T_\theta > 0\Big) = \frac{\alpha}{M}. \tag{3.5}$$

Because T_{ABT} and T_θ have the same asymptotic distribution and they are both independent of stage 1 analysis, $c_2^* = c_2$. The asymptotic power using J_θ is similar to (3.5), but evaluated under H_1 (see Appendix B).

3.3. Using the ABT in stage 1 and GMS in stage 2

In Section 3.2, the genetic model is assumed to be known. For many common and complex diseases, however, the genetic models are usually unknown to the researchers. In this case, J_θ cannot be directly applied without specifying θ. In practice, $J_{1/2}$ or J_{ABT} may be applied as a robust choice regardless of the true genetic model. Here we apply the GMS (Zheng and Ng (2008)) in the second stage.

The two-stage GMS method works as follows. If $T_{\mathrm{pool}} > 0$, then allele A is regarded as the risk allele and we set $T_{\mathrm{model}} = T_0$ if $T_{\mathrm{HWD}} > c_0$, $T_{\mathrm{model}} = T_1$ if $T_{\mathrm{HWD}} < c_0$, and $T_{\mathrm{model}} = T_{1/2}$ if $|T_{\mathrm{HWD}}| \le c_0$, where $c_0 = 1.645$ as in Zheng and Ng (2008). On the other hand, if $T_{\mathrm{pool}} < 0$, then we can switch alleles A and a and apply the above GMS similarly. The joint analysis is written as

$$J_{\mathrm{GMS}} = \omega^{1/2} T_{\mathrm{pool}} + (1-\omega)^{1/2} T_{\mathrm{model}}. \tag{3.6}$$

Note that in Zheng and Ng (2008), the risk allele is also the minor allele or it is known. In our two-stage design, the risk allele is determined in stage 1. Thus, we do not need to know the risk allele or to use the minor allele as the risk allele. This is one advantage of the two-stage analysis. In the second stage, the information about the risk allele is free because the Type I error for determining the risk allele has been paid in the first stage. In fact, by the symmetry of the normal distribution, it can be shown that the above procedure has the same asymptotic Type I error.

To apply the joint analysis J_{GMS} the threshold value c_1 is given as before, and c_2^{**} for stage 2 is determined by

$$P_{H_0}\Big(|T_{\mathrm{pool}}| > c_1,\ |J_{\mathrm{GMS}}| > c_2^{**},\ T_{\mathrm{model}} \cdot T_{\mathrm{pool}} > 0\Big) = \frac{\alpha}{M}. \tag{3.7}$$

The asymptotic power for the joint analysis J_{GMS} can be obtained from (3.7) evaluated under H_1 (see Appendix C).

4. Results

4.1. Simulation studies

Three joint analysis strategies (J_{ABT}, J_θ and J_{GMS}) for the two-stage design with DNA pooling have been discussed in Section 3. They all have DNA pooling with the ABT in the first stage, but have different procedures (the ABT, optimal

CATT and GMS) in the second stage. In the following, we refer to these three approaches as procedures II-ABT, II-CATT and II-GMS. Zuo et al. (2006) and Ji et al. (2007) showed that the two-stage design is often more powerful with equal fraction of samples in the two stages. Thus, we conducted simulation studies using 1,000 cases and 1,000 controls that were split for the two stages with equal proportion ($r = s = r_* = s_* = 500$). We also conducted simulations using smaller sample size and got similar results (results are not reported here).

Four common genetic models were considered: REC, ADD, MUL and DOM. For each model we set GRR (λ_2) at 1.5, 1.8, 2.0 and 2.5 and the risk allele frequency in the population to be $p = 0.1$, 0.3 and 0.5. Our GRR was taken to be much smaller than that in Zuo et al. (2006) in which GRR was taken to be 4.0 under various models. The measurement error was assumed to be fixed at $\epsilon^2 = 0$, 0.005, 0.01 and 0.03. We considered two DNA pooling settings: a single pool ($m = 1$) and four pools ($m = 4$), similar to those used in Zuo et al. (2006). Note that Zuo et al. (2006) only presented numerical results with a single pool. The genome-wide level for testing 300,000 SNPs is 0.05, so the Type I error for a single SNP was 1.67×10^{-7} by the Bonferroni correction. After the DNA pooling, the top 5% ($\alpha_1 = 0.05$) SNPs were selected for stage 2. Zuo et al. (2006) and Gail, Pfeiffer, Wheeler and Pee (2007) both suggested choosing the top 5% for genome-wide scans. Given the above settings, our numerical results showed that the threshold values are $c_1 = 1.96$, $c_2 = c_2^* = 5.232$, and $c_2^{**} = 5.308$ (5.319, 5.323) when the minor allele frequency $p = 0.1$ (0.3, 0.5), where only c_2^{**} depends on the allele frequency. In each setting, results were obtained based on 100,000 replicates. We estimated the power for the above three procedures and report relative power ratios under different parametric settings.

4.2. Comparing procedures II-ABT and II-CATT

Table 1 reports the power comparison between J_θ and J_{ABT} when the genetic model is known (either recessive or dominant). We define the relative efficiency (RE) as the ratio of the empirical power of II-CATT over that of II-ABT. When $m = 1$ and $\epsilon^2 = 0$, there is no difference between DNA pooling and individual genotyping in estimating the allele frequency. Thus, the RE is equal to that of the comparison between using the ABT and the optimal CATT based on the joint analysis of Skol et al. (2006). When the underlying genetic model was REC or DOM (Table 1), using the second design was always more powerful than using the first design. The gain in power could be substantial (RE is up to 3.5) for common allele p and moderately large GRR. The gain also increased with ϵ, which indicates that the design with optimal CATTs in stage 2 is more robust to the measurement errors under the REC and DOM models. The gain in power under these two models is not surprising because the optimal CATTs are used for the

Table 1. Relative efficiency (RE) of joint analysis in stage 2 (RE = empirical power of the optimal CATT over the power of the ABT) when DNA pooling is employed in stage 1: REC and DOM models.

			ϵ^2							
			$m=1$				$m=4$			
Model	λ_2	p	0	0.005	0.01	0.03	0	0.005	0.01	0.03
REC	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	*	*	*	*	*	*	*	*
		0.5	1.63	1.64	1.73	2.27	1.59	1.64	1.67	1.89
	1.8	0.1	*	*	*	*	*	*	*	*
		0.3	2.63	2.64	2.75	3.35	2.60	2.62	2.67	3.19
		0.5	1.17	1.19	1.26	1.54	1.16	1.16	1.18	1.36
	2.0	0.1	*	*	*	*	*	*	*	*
		0.3	2.10	2.18	2.43	3.49	2.08	2.09	2.25	2.61
		0.5	1.03	1.03	1.06	1.28	1.03	1.03	1.03	1.12
	2.5	0.1	*	*	*	*	*	*	*	*
		0.3	1.16	1.20	1.32	1.87	1.16	1.17	1.20	1.47
		0.5	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
DOM	1.5	0.1	1.13	1.14	1.23	1.25	1.15	1.15	1.15	1.21
		0.3	1.45	1.47	1.56	1.80	1.48	1.50	1.53	1.62
		0.5	2.38	2.85	3.27	3.47	2.55	2.77	2.80	3.33
	1.8	0.1	1.04	1.06	1.10	1.21	1.03	1.05	1.06	1.13
		0.3	1.09	1.12	1.20	1.54	1.10	1.10	1.13	1.27
		0.5	2.18	2.30	2.56	4.15	2.14	2.26	2.30	2.84
	2.0	0.1	1.01	1.01	1.03	1.13	1.01	1.01	1.01	1.07
		0.3	1.01	1.02	1.06	1.28	1.01	1.02	1.02	1.11
		0.5	1.71	1.82	2.07	3.30	1.74	1.76	1.82	2.38
	2.5	0.1	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
		0.3	1.00	1.00	1.00	1.02	1.00	1.00	1.00	1.00
		0.5	1.15	1.18	1.32	2.06	1.15	1.16	1.18	1.52

* The powers of the ABT and the CATT are approximately 0.

REC and DOM models (Sasieni (1997) and Freidlin et al. (2002)). The results for the ADD and MUL models are reported in Table 2. From Sasieni (1997) and Zheng et al. (2003), the ABT and the additive CATT are asymptotically equivalent. Thus, the REs in Table 2 are all close to 1 under the ADD model. For the MUL model, the ABT seems to be slightly more powerful than the additive CATT . The REs in Table 2 do not change noticeably with the measurement errors ϵ. To summarize, when the underlying genetic models are known, using optimal CATT was preferable to using the ABT in the second stage.

4.3. Comparing II-ABT, II-CATT and II-GMS

To examine the performance of II-GMS, we first compared it with II-CATT under the REC and DOM models even when the underlying models were known.

Table 2. Relative efficiency (RE) of joint analysis in stage 2 (RE = empirical power of the optimal CATT over the power of the ABT) when DNA pooling is employed in stage 1: ADD and MUL models.

			ϵ^2							
			$m=1$				$m=4$			
Model	λ_2	p	0	0.005	0.01	0.03	0	0.005	0.01	0.03
ADD	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.97	0.99	1.00	1.00	0.98	0.95	0.98	0.97
		0.5	0.98	0.99	0.94	0.93	0.98	0.97	0.98	0.98
	1.8	0.1	1.00	0.98	0.99	1.00	1.01	0.97	0.95	1.03
		0.3	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98
		0.5	1.00	0.99	0.98	0.98	1.00	0.99	0.99	0.99
	2.0	0.1	0.99	1.00	0.96	0.98	0.99	0.99	0.99	0.95
		0.3	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99
		0.5	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
	2.5	0.1	1.00	1.00	0.99	0.97	1.00	1.00	1.00	1.00
		0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.5	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
MUL	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.97	0.99	0.98	0.84	0.98	0.97	0.96	0.99
		0.5	0.95	0.92	0.94	0.93	0.97	0.97	0.97	0.95
	1.8	0.1	0.96	0.95	0.94	0.94	0.98	0.97	0.95	0.96
		0.3	0.98	0.98	0.98	0.92	0.98	0.99	0.98	0.97
		0.5	0.98	0.98	0.98	0.96	0.99	0.99	0.98	0.98
	2.0	0.1	0.98	0.98	0.96	0.89	0.98	0.98	0.97	0.95
		0.3	0.99	0.99	0.98	0.96	0.99	0.99	0.99	0.98
		0.5	1.00	0.99	0.99	0.96	1.00	0.99	0.99	0.98
	2.5	0.1	0.99	0.98	0.96	0.98	0.99	0.98	0.98	0.97
		0.3	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
		0.5	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00

* The powers of the ABT and the CATT are approximately 0.

Results are reported in Table 3. Note that II-GMS performed reasonably well compared to II-CATT for the given models. Most REs were greater than 0.85, with one RE less than 0.80.

For genome-wide association studies, however, the underlying genetic models of SNPs with true association are usually unknown. Thus, we propose to use the joint analysis using GMS for two-stage design with DNA pooling. We compare the REs, defined as before, of II-ABT with II-GMS in stage 2. Results for the REC and DOM models are reported in Table 4 and for the ADD and MUL models in Table 5. From Table 4, II-GMS was overall more powerful than II-ABT. Similar to Tables 1 and 2, II-GMS could gain substantial power compared to II-ABT. The gain in power also increased with the measurement errors ϵ. On

Table 3. Relative efficiency (RE) of joint analysis in stage 2 (RE = empirical power of the GMS over the power of the optimal CATT) when DNA pooling is employed in stage 1: REC and DOM models. The underlying genetic model is known.

			ϵ^2							
			$m=1$				$m=4$			
Model	λ_2	p	0	0.005	0.01	0.03	0	0.005	0.01	0.03
REC	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	*	*	*	*	*	*	*	*
		0.5	0.88	0.88	0.86	0.83	0.88	0.89	0.86	0.85
	1.8	0.1	*	*	*	*	*	*	*	*
		0.3	0.89	0.89	0.87	0.86	0.89	0.89	0.88	0.86
		0.5	0.97	0.97	0.96	0.93	0.98	0.98	0.97	0.95
	2.0	0.1	*	*	*	*	*	*	*	*
		0.3	0.94	0.93	0.92	0.88	0.93	0.93	0.92	0.89
		0.5	1.00	1.00	0.99	0.96	1.00	1.00	1.00	0.98
	2.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.99	0.99	0.98	0.95	0.99	0.99	0.99	0.98
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DOM	1.5	0.1	0.88	0.91	0.86	0.83	0.88	0.89	0.90	0.88
		0.3	0.91	0.91	0.89	0.87	0.91	0.91	0.92	0.89
		0.5	0.89	0.83	0.86	0.76	0.88	0.85	0.86	0.81
	1.8	0.1	0.98	0.97	0.95	0.91	0.98	0.97	0.97	0.93
		0.3	0.99	0.98	0.98	0.94	0.99	0.99	0.98	0.97
		0.5	0.93	0.92	0.91	0.85	0.92	0.93	0.93	0.89
	2.0	0.1	1.00	0.99	0.98	0.95	1.00	1.00	0.99	0.97
		0.3	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.99
		0.5	0.96	0.96	0.95	0.91	0.96	0.96	0.96	0.93
	2.5	0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.5	1.00	0.99	0.99	0.97	1.00	1.00	0.99	0.99

* The powers of the GMS and the CATT are approximately 0.

the other hand, in Table 5, since the ABT is asymptotically equivalent to the additive CATT, II-GMS was less powerful compared to II-ABT under the ADD or MUL models in the two-stage design. However, the loss of power from using II-GMS was slight in most situations, although the power loss increased with ϵ and decreased with λ_2. Under the REC model, when GRR = 1.8 and $p = 0.3$, the RE was about 2.5 using II-GMS compared to using II-ABT. For the DOM model, II-GMS and II-ABT had similar power except for the common allele frequencies, under which, e.g., the RE was about 2 when GRR = 1.8 and $p = 0.5$. For the ADD and MUL models, the largest loss of the power using II-GMS occured when GRR = 1.5. When GRR = 1.8 , the RE using II-GMS was greater than 0.8 for $p = 0.1$, and greater than 0.90 for $p = 0.3$. Thus, based on the results of the four genetic models, II-GMS was more robust than II-ABT in the sense that it

Table 4. Relative efficiency (RE) of joint analysis in stage 2 (RE = empirical power of the GMS over the power of the ABT) when DNA pooling is employed in stage 1: REC and DOM models. The underlying genetic model is unknown.

			ϵ^2							
			$m=1$				$m=4$			
Model	λ_2	p	0	0.005	0.01	0.03	0	0.005	0.01	0.03
REC	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	*	*	*	*	*	*	*	*
		0.5	0.88	0.88	0.86	0.83	0.88	0.89	0.86	0.85
	1.8	0.1	*	*	*	*	*	*	*	*
		0.3	0.89	0.89	0.87	0.86	0.89	0.89	0.88	0.86
		0.5	0.97	0.97	0.96	0.93	0.98	0.98	0.97	0.95
	2.0	0.1	*	*	*	*	*	*	*	*
		0.3	0.94	0.93	0.92	0.88	0.93	0.93	0.92	0.89
		0.5	1.00	1.00	0.99	0.96	1.00	1.00	1.00	0.98
	2.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.99	0.99	0.98	0.95	0.99	0.99	0.99	0.98
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DOM	1.5	0.1	0.88	0.91	0.86	0.83	0.88	0.89	0.90	0.88
		0.3	0.91	0.91	0.89	0.87	0.91	0.91	0.92	0.89
		0.5	0.89	0.83	0.86	0.76	0.88	0.85	0.86	0.81
	1.8	0.1	0.98	0.97	0.95	0.91	0.98	0.97	0.97	0.93
		0.3	0.99	0.98	0.98	0.94	0.99	0.99	0.98	0.97
		0.5	0.93	0.92	0.91	0.85	0.92	0.93	0.93	0.89
	2.0	0.1	1.00	0.99	0.98	0.95	1.00	1.00	0.99	0.97
		0.3	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.99
		0.5	0.96	0.96	0.95	0.91	0.96	0.96	0.96	0.93
	2.5	0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.5	1.00	0.99	0.99	0.97	1.00	1.00	0.99	0.99

* The powers of the GMS and the CATT are approximately 0.

suffered minor power loss under the ADD/MUL models, relative to more gains in power under the REC/DOM models.

Acknowledgement

The work of Yuan was completed while she was visiting Department of Statistics, Columbia University, partially supported by the Chinese Council of Scholarship. She also thanks Dr. Zhiliang Ying for his encouragement and support during her visit. The work of Yuan and Yang was also partially supported by a China NSF Grant and a grant from the Chinese Academy of Science. We would like to thank two reviewers for their insightful suggestions and comments which

Table 5. Relative efficiency (RE) of joint analysis in stage 2 (RE = empirical power of the GMS over the power of the ABT) when DNA pooling is employed in stage 1: ADD and MUL models. The underlying genetic model is unknown.

			ϵ^2							
			$m=1$				$m=4$			
Model	λ_2	p	0	0.005	0.01	0.03	0	0.005	0.01	0.03
ADD	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.90	0.90	0.91	1.00	0.88	0.95	0.93	0.85
		0.5	0.89	0.90	0.88	0.83	0.89	0.95	0.90	0.93
	1.8	0.1	0.85	0.86	0.84	0.86	0.92	0.87	0.84	0.88
		0.3	0.96	0.96	0.95	0.89	0.96	0.97	0.96	0.92
		0.5	0.95	0.94	0.92	0.88	0.96	0.96	0.95	0.92
	2.0	0.1	0.92	0.92	0.87	0.86	0.92	0.91	0.89	0.83
		0.3	0.99	0.98	0.97	0.93	0.98	0.98	0.98	0.96
		0.5	0.98	0.98	0.97	0.92	0.98	0.97	0.97	0.96
	2.5	0.1	0.98	0.97	0.96	0.91	0.98	0.98	0.97	0.95
		0.3	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
		0.5	1.00	1.00	0.99	0.98	1.00	1.00	1.00	0.99
MUL	1.5	0.1	*	*	*	*	*	*	*	*
		0.3	0.85	0.90	1.00	0.79	0.88	0.96	0.88	0.84
		0.5	0.89	0.86	0.84	0.80	0.89	0.90	0.92	0.84
	1.8	0.1	0.84	0.89	0.73	0.81	0.86	0.81	0.86	0.86
		0.3	0.93	0.92	0.90	0.82	0.93	0.93	0.93	0.87
		0.5	0.94	0.94	0.93	0.84	0.95	0.95	0.94	0.93
	2.0	0.1	0.86	0.86	0.83	0.70	0.86	0.88	0.83	0.80
		0.3	0.96	0.95	0.94	0.86	0.96	0.96	0.96	0.92
		0.5	0.98	0.96	0.95	0.90	0.97	0.98	0.97	0.93
	2.5	0.1	0.93	0.91	0.87	0.82	0.93	0.93	0.92	0.84
		0.3	1.00	1.00	0.99	0.95	1.00	1.00	1.00	0.98
		0.5	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.99

* The powers of the GMS and ABT are approximately 0.

have greatly improved our presentation.

Appendix A

Let x_{ij} (y_{ij}) be the number of allele A carried by the jth individual in the ith pool in cases (controls), and u_i and v_i be the i.i.d. disturbance variables from $N(0,\epsilon^2)$. Let $\hat{p}_{1i}^{\text{pool}} = 2h_1^{-1}\sum_{j=1}^{h_1} x_{ij} + u_i$ and $\hat{p}_{0i}^{\text{pool}} = 2h_0^{-1}\sum_{j=1}^{h_0} y_{ij} + v_i$. Then write $\hat{p}_1^{\text{pool}} = m^{-1}\sum_{i=1}^{m}\hat{p}_{1i}^{\text{pool}}$, $\hat{p}_0^{\text{pool}} = m^{-1}\sum_{i=1}^{m}\hat{p}_{0i}^{\text{pool}}$ and $\hat{p}^{\text{pool}} = \psi\hat{p}_1^{\text{pool}} + (1-\psi)\hat{p}_0^{\text{pool}}$, where $\psi = r/n$.

Note that c_1 is the $100(1-\alpha_1/2)$th percentile of $N(0,1)$. For c_2, under H_0,

T_{pool} and T_{ABT} are independent and asymptotically $N(0,1)$, distribution $\Phi(x)$ and density $\phi(x)$. Thus, c_2 asymptotically satisfies

$$\iint_{R_1} \phi(x)\phi(y)dxdy = 2\iint_{R_2} \phi(x)\phi(y)dxdy = \frac{\alpha}{M}, \tag{A.1}$$

where $R_1 = \{|x| > c_1,\ |w_1x + w_2y| > c_2,\ xy > 0\}$, $R_2 = \{x > c_1,\ w_1x + w_2y > c_2, y > 0\}$, $w_1 = \omega^{1/2}$, and $w_2 = (1-\omega)^{1/2}$. Further, (A.1) can be written as

$$\int_{c_1}^{c_2/w_1} \Phi\Big(\frac{n^{1/2}x - (n+n_*)^{1/2}c_2}{n_*^{1/2}}\Big)d\Phi(x) + \frac{1}{2}\Phi\Big(-\frac{c_2}{w_1}\Big) = \frac{\alpha}{2M},$$

from which c_2 can be solved numerically.

In order to calculate the asymptotic power for J_{ABT} for a given genetic model with the joint distribution of T_{pool} and T_{ABT} under H_1, we need to compute the means and variances of the two statistics under H_1. Write $p_1^{\text{pool}} = p_2 + p_1/2$ and $p_0^{\text{pool}} = q_2 + q_1/2$, with estimates given before. Since (u_i, v_i) are independent of genotypes, $\mu = E_{H_1}(\hat{p}_1^{\text{pool}} - \hat{p}_0^{\text{pool}}) = p_1^{\text{pool}} - p_0^{\text{pool}}$ and $\text{Var}_{H_1}(\hat{p}_1^{\text{pool}} - \hat{p}_0^{\text{pool}}) = \sigma^{*2} + 2\epsilon^2/m$, where $\sigma^{*2} = \{4p_2 + p_1 - (2p_2+p_1)^2\}/(4r) + \{4q_2 + q_1 - (2q_2 + q_1)^2\}/(4s)$. Let $p^* = \psi p_1^{\text{pool}} + (1-\psi)p_0^{\text{pool}}$. Then $E_{H_1}(\hat{p}^{\text{pool}}) = p^*$. Define $\sigma^2 = p^*(1-p^*)\{1/(2r) + 1/(2s)\}$. Let $Z_1 \sim N(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\mu}{(\sigma^2 + 2\epsilon^2/m)^{1/2}}, \quad \sigma_1^2 = \frac{\sigma^{*2} + 2\epsilon^2/m}{\sigma^2 + 2\epsilon^2/m}.$$

Then, under H_1, T_{pool} and Z_1 have the same asymptotic distribution.

For stage 2, let $\psi_* = r_*/n_*$ and p_{case} with p_{cont} used to denote the allele A's frequencies in case and control groups, $p_* = \psi_* p_{\text{case}} + (1-\psi_*)p_{\text{cont}} = E_{H_1}(\hat{p})$, and $\hat{p}$, given in T_{ABT}, is the allele frequency estimate from data in stage 2 under the null. Write $\sigma_*^2 = p_*(1-p_*)\{1/(2r_*) + 1/(2s_*)\}$. Similar to the above derivations for stage 1, for stage 2, we have asymptotically that T_{ABT} and Z_2 have the same asymptotic distribution where $Z_2 \sim N(\mu_2, \sigma_2^2)$ under H_1, with $\mu_2 = \mu/\sigma_*$ and $\sigma_2^2 = \sigma^{*2}/\sigma_*^2$. Let $\Phi_i(x)$ be the distribution function of $N(\mu_i, \sigma_i^2)$ for $i = 1, 2$, then the asymptotic power of J_{ABT}, π_{ABT}, is

$$\pi_{\text{ABT}} = \iint_{R_1} d\Phi_1(x)d\Phi_2(y) = \iint_{R_2} d\Phi_1(x)d\Phi_2(y) + \iint_{R_3} d\Phi_1(x)d\Phi_2(y),$$

where $R_3 = \{x < -c_1,\ w_1x + w_2y < -c_2,\ y < 0\}$.

Appendix B

The asymptotic power of J_θ, π_{CATT}, is similar to π_{ABT} with T_{ABT} being replaced by T_θ. The correlations among the test statistics are different under H_1.

In the following, the higher order terms are omitted. Let $U_\theta = p_2+\theta p_1 - q_2 - \theta q_1$ and $\hat{U}_\theta = \hat{p}_2 + \theta\hat{p}_1 - \hat{q}_2 - \theta\hat{q}_1$, where $\hat{p}_l = r_{*l}/r_*$ and $\hat{q}_l = s_{*l}/s_*$ for $l = 0, 1, 2$. (Note that $\hat{p}_l = r_l/r$ and $\hat{q}_l = s_l/s$ were used before.) Under H_0, $p_l = q_l = \pi_l$ for $l = 0, 1, 2$. Then $E_{H_0}(\hat{U}_\theta) = 0$, $\mathrm{Var}_{H_0}(\hat{U}_\theta) = \{\pi_2+\theta^2\pi_1-(\pi_2+\theta\pi_1)^2\}(1/r_*+1/s_*)$, which can be estimated by $\widehat{\mathrm{Var}}_{H_0}(\hat{U}_\theta)$, where $\hat{\pi}_l = (r_{*l} + s_{*l})/n_* = n_{*l}/n_*$. Write $\sigma^{2*}_\theta = n_*\widehat{\mathrm{Var}}_{H_0}(\hat{U}_\theta)$, so T_θ can be written as $T_\theta = n_*^{1/2}\hat{U}_\theta/\sigma^*_\theta$ in distribution. Let $\mu_\theta = E_{H_1}(\hat{U}_\theta)$ and $\sigma^2_\theta = \mathrm{Var}_{H_1}(n_*^{1/2}\hat{U}_\theta) = \{(\theta^2 p_1 + p_2 - (\theta p_1 + p_2)^2)\}(n_*/r_*) + \{(\theta^2 q_1 + q_2 - (\theta q_1 + q_2)^2)\}(n_*/s_*)$. Then, under H_1, T_θ and Z_3 have the same asymptotic distribution, where $Z_3 \sim N(\mu_3, \sigma_3^2)$ with distribution function $\Phi_3(x)$, where $\mu_3 = n_*^{1/2}\mu_\theta/\sigma^*_\theta$ and $\sigma_3^2 = \sigma^2_\theta/\sigma^{2*}_\theta$. Then the asymptotic power can be written as

$$\pi_{\mathrm{CATT}} = \iint_{R_1} d\Phi_1(x)d\Phi_3(y) = \iint_{R_2} d\Phi_1(x)d\Phi_3(y) + \iint_{R_3} d\Phi_1(x)d\Phi_3(y).$$

Appendix C

Write under either H_0 or H_1,

$$\Sigma_1 = \mathrm{Var}\begin{pmatrix}\hat{p}_1\\ \hat{p}_2\end{pmatrix} = \frac{1}{r_*}\begin{pmatrix} p_1(1-p_1) & -p_1p_2\\ -p_1p_2 & p_2(1-p_2)\end{pmatrix},$$

$$\Sigma_0 = \mathrm{Var}\begin{pmatrix}\hat{q}_1\\ \hat{q}_2\end{pmatrix} = \frac{1}{s_*}\begin{pmatrix} q_1(1-q_1) & -q_1q_2\\ -q_1q_2 & q_2(1-q_2)\end{pmatrix}.$$

Let $f(x,y) = x(1-x)(x/2+y)^2+2xy(x/2+y)(1-x-2y)+y(1-y)(1-x-2y)^2$. Under H_0, $E_{H_0}(\hat{\delta}_1 - \hat{\delta}_0) = 0$ and $\mathrm{Var}_{H_0}(\hat{\delta}_1 - \hat{\delta}_0) = f(\pi_1, \pi_2)(1/r_* + 1/s_*)$. If $\sigma^{2*}_{\mathrm{HWD}} = n_*\widehat{\mathrm{Var}}_{H_0}(\hat{\delta}_1 - \hat{\delta}_0) = f(\hat{\pi}_1, \hat{\pi}_2)(n_*/r_* + n_*/s_*)$, we can write T_{HWD} as

$$T_{\mathrm{HWD}} = \frac{n_*^{1/2}(\hat{\delta}_1 - \hat{\delta}_0)}{\sigma^*_{\mathrm{HWD}}}.$$

We can write $\sigma^2_{\mathrm{HWD}} = n_*\mathrm{Var}_{H_1}(\hat{\delta}_1 - \hat{\delta}_0) = f(p_1, p_2)(n_*/r_*) + f(q_1, q_2)(n_*/s_*)$ by the Delta method. Therefore, T_{HWD} and Z_4 have the same asymptotic distribution, where $Z_4 \sim N(\mu_4, \sigma_4^2)$ under H_1 with $\mu_4 = n_*^{1/2}(\delta_1 - \delta_0)/\sigma^*_{\mathrm{HWD}}$ and $\sigma_4^2 = \sigma^2_{\mathrm{HWD}}/\sigma^{2*}_{\mathrm{HWD}}$.

Let $T_{\mathrm{pool}} = x$. To find the threshold c_2^{**} in (3.7), the left hand side of (3.7) can be written as

$$\begin{aligned}
&P_{H_0}(x > c_1, T_{\mathrm{HWD}} > c_0, w_1x + w_2T_0 > c_2^{**}, T_0 > 0)\\
&\quad + P_{H_0}(x > c_1, T_{\mathrm{HWD}} < -c_0, w_1x + w_2T_1 > c_2^{**}, T_1 > 0)\\
&\quad + P_{H_0}(x < -c_1, T_{\mathrm{HWD}} > c_0, w_1x + w_2T_1 < -c_2^{**}, T_1 < 0)\\
&\quad + P_{H_0}(x < -c_1, T_{\mathrm{HWD}} < -c_0, w_1x + w_2T_0 < -c_2^{**}, T_0 < 0)\\
&\quad + P_{H_0}(|x| > c_1, |T_{\mathrm{HWD}}| \le c_0, |w_1x + w_2T_{1/2}| > c_2^{**}, T_{1/2}\cdot x > 0), \quad \text{(A.2)}
\end{aligned}$$

where each probability is a function of the correlation between T_{HWD} and T_θ in the second stage. From Zheng and Ng (2008), $\mathrm{corr}_{H_0}(T_{\mathrm{HWD}}, T_0) = \{(1-p)/(1+p)\}^{1/2} + O(n^{-1})$, $\mathrm{corr}_{H_0}(T_{\mathrm{HWD}}, T_1) = -\{p/(2-p)\}^{1/2} + O(n^{-1})$, and T_{HWD} and $T_{1/2}$ are asymptotically independent under H_0 with order $O(n^{-1})$. Let $\rho_0 = \{(1-p)/(1+p)\}^{1/2}$ and $\rho_1 = -\{p/(2-p)\}^{1/2}$. Let $\Phi^0(y,z)$ and $\Phi^1(y,z)$ be the distribution of bivariate normal with mean $(0,0)$ and covariance matrices $\Lambda_0 = \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$ and $\Lambda_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}$, respectively. Let $A_1 = \{x > c_1, y > c_0, z > 0, w_1x + w_2z > c_2^{**}\}$, $A_2 = \{x > c_1, y < -c_0, z > 0, w_1x + w_2z > c_2^{**}\}$, $A_3 = \{x < -c_1, y > c_0, z < 0, w_1x + w_2z < -c_2^{**}\}$, $A_4 = \{x < -c_1, y < -c_0, z < 0, w_1x + w_2z < -c_2^{**}\}$, and $A_5 = \{|x| > c_1, |y| \le c_0, z < 0, |w_1x + w_2z| > c_2^{**}\}$. Then (A.2) can be written as $\int_{A_1} d\Phi(x)d\Phi^0(y,z) + \int_{A_2} d\Phi(x)d\Phi^1(y,z) + \int_{A_3} d\Phi(x)d\Phi^1(y,z) + \int_{A_4} d\Phi(x)d\Phi^0(y,z) + \int_{A_5} d\Phi(x)d\Phi(y)d\Phi(z)$.

To obtain the power of J_{GMS}, we need the correlation between T_θ and T_{HWD} under H_1,

$$\rho_\theta^* = \mathrm{corr}_{H_1}(T_{\mathrm{HWD}}, T_\theta) = \frac{n_*}{\sigma_{\mathrm{HWD}}\sigma_\theta}\left\{\mathrm{Cov}_{H_1}\left(\hat{\delta}_1, \hat{U}_\theta\right) - \mathrm{Cov}_{H_1}\left(\hat{\delta}_0, \hat{U}_\theta\right)\right\}.$$

Let $f_1(x,y) = y - (y + x/2)^2$ and $f_2(x,y) = y + \theta x$. Then

$$\mathrm{Cov}_{H_1}\left(\hat{\delta}_1, \hat{U}_\theta\right) = \mathrm{Cov}_{H_1}\left(f_1(\hat{p_1}, \hat{p_2}), f_2(\hat{p_1}, \hat{p_2})\right).$$

A similar expression can be obtained for $\mathrm{Cov}_{H_1}(\hat{\delta}_0, \hat{U}_\theta)$. Using the Delta method,

$$\mathrm{Cov}_{H_1}\left(\hat{\delta}_1, \hat{U}_\theta\right) = \left(\frac{\partial f_1(p_1,p_2)}{\partial p_1} \quad \frac{\partial f_1(p_1,p_2)}{\partial p_2}\right)\Sigma_1\begin{pmatrix} \frac{\partial f_2(p_1,p_2)}{\partial p_1} \\ \frac{\partial f_2(p_1,p_2)}{\partial p_2} \end{pmatrix}. \tag{A.3}$$

Write $g_\theta(x,y) = \theta\{x(1-x)(y+x/2) + xy(1-x-2y)\} - \{xy(y+x/2) + y(1-y)(1-x-2y)\}$ and $\phi_* = \lim r_*/n_*$.

Let $\xi_\theta^* = -\{g_\theta(p_1,p_2)/\phi_* + g_\theta(q_1,q_2)/(1-\phi_*)\}/(\sigma_{\mathrm{hwd}}\sigma_\theta)$. Under H_1, Z_5 has a bivariate normal distribution with mean vector $(\mu_4, \mu_3)'$ and covariance matrix

$$\begin{pmatrix} \dfrac{\sigma_{\mathrm{hwd}}^2}{\sigma_{\mathrm{hwd}}^{2*}} & -\xi_\theta^*\dfrac{\sigma_{\mathrm{hwd}}\sigma_\theta}{\sigma_{\mathrm{hwd}}^*\sigma_\theta^*} \\ -\xi_\theta^*\dfrac{\sigma_{\mathrm{hwd}}\sigma_\theta}{\sigma_{\mathrm{hwd}}^*\sigma_\theta^*} & \dfrac{\sigma_\theta^2}{\sigma_\theta^{2*}} \end{pmatrix}. \tag{A.4}$$

Then $(T_{\mathrm{hwd}}, T_\theta)'$ and Z_5 have same asymptotic distribution under H_1. If $\tilde{\Phi}^\theta(x,y)$ is the joint distribution function of T_{HWD} and T_θ, the power function of J_{GMS}

can be written as

$$\pi_{\mathrm{GMS}} = \int_{A_1} d\Phi(x) d\tilde{\Phi}^0(y,z) + \int_{A_2} d\Phi(x) d\tilde{\Phi}^1(y,z) + \int_{A_3} d\Phi(x) d\tilde{\Phi}^1(y,z) + \int_{A_4} d\Phi(x) d\tilde{\Phi}^0(y,z) + \int_{A_5} d\Phi(x) d\tilde{\Phi}^{1/2}(y,z). \tag{A.5}$$

Regions A_i, $i = 1, \ldots, 5$ are given as above.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-386.

Barcellos, L. F., Klitz, W., Field, L. L., Tobias, R., Bowcock, A. M., Wilson, R., Nelson, M. P., Nagatomi, J. and Thomson, G. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61**, 734-747.

Barratt, B. J., Payne, F., Rance, H. E., Nutland, S., Todd, J. A. and Clayton, D. G. (2002). Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.* **66**, 393-405.

Bukszar, J. and van den Oord, E. (2006). Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for Pearson's statistic. *Biometrics* **62**, 1132-1137.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* **10**, 417-451.

Dube, M. P., Schmidt, S. and Hauser, E. (2007). Multistage designs in the genomic era: Providing balance in complex disease studies. *Genet. Epidemiol.* **S31**, S118.

Elston, R. C., Lin, D. Y. and Zheng, G. (2007). Multistage sampling for genetic studies. *Ann. Rev. Genomics Hum. Genet.* **8**, 327-342.

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

Gail, M. H., Pfeiffer, R. M., Wheeler, W. and Pee, D. (2007). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* **9**, 201-215.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4**, 701-709.

Ji, F., Stephen, J. F., Chad, H., Nancy, R. M. and Derek, G. (2007). Incorporation of genetic model parameters for cost-effective designs of genetic association studies using DNA pooling. *BMC Genomics* **8**, 238.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.* **78**, 505-509.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413-417.

Norton, N., Williams, N. M., O'Donovan, M. C. and Owen, M. J. (2004). DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.* **36**, 146-52.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics* **53**, 1253-1261.

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58**, 163-70.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25**, 149-157.

Satagopan, J. M., Venkatraman, E. S. and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589-597.

Schaid, D. J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M. and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human diseases. *Am. J. Hum. Genet.* **76**, 780-793.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M. and Owen, M. (2002). DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **11**, 862-871.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209-213.

Song, K. and Elston, R. C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat. Med.* **25**, 105-126.

Suh, Y. J. and Li, W. (2007). Genotype-based case-control analysis, violation of Hardy-Weinberg equilibrium, and phase diagram. *Proceedings of the 5th Asia-Pacific Bioinformatics Conference*, 185-194. Imperial College Press.

Thomas, D., Xie, R. R. and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* **27**, 401-414.

Thomas, D. C., Haile, R. W. and Duggan, D. (2005). Recent developments in genomewide association scans: A workshop summary and review. *Am. J. Hum. Genet.* **77**, 337-345.

Wang, H. S., Thomas, D. C., Pe'er, I. and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* **30**, 356-368.

Wang, T., Zhu, X. and Elston, R. C. (2007). Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am. J. Hum. Genet.* **80**, 911-920.

Wittke-Thompson, J. K., Pluzhnikov, A. and Cox, N. J. (2005). Rational inferences about departure from Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967-986.

Zaykin, D. V. and Nielsen, D. M. (2000). Hardy-Weinberg disequilibrium (HWD) fine mapping for case-control samples. *Am. J. Hum. Genet.* **67**, 1238 Suppl.

Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical J.* **45**, 335-348.

Zheng, G. and Ng, H. K. T. (2008). Genetic model selection in two-stage analysis for case-control association studies. *Biostatistics* **9**, 391-399.

Zuo, Y. J, Zou, G. H. and Zhao, H. Y. (2006). Two-stage designs in case-control association analysis. *Genetics* **173**, 1747-1760.

Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui, China.

E-mail: myuan@mail.ustc.edu.cn

Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui, China.

E-mail: ynyang@ustc.edu.cn

Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892, U.S.A.

E-mail: zhengg@nhlbi.nih.gov

(Received January 2008; accepted September 2008)